USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA
FACULTADE DE MATEMÁTICAS

Traballo Fin de Grao

REGRESIÓN LINEAL CON DATOS CENSURADOS

María Barreira Miranda

2019/2020

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

GRAO DE MATEMÁTICAS

Traballo Fin de Grao

Regresión lineal con datos censurados

María Barreira Miranda

Xullo 2020

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Traballo proposto

Área de Coñecemento: Estatística e Investigación Operativa
Título: Regresión lineal con datos censurados
Breve descrición do contido
Os datos censurados son moi habituais na Análise de Supervivencia, que é a parte da Estatística que estuda os tempos de vida. E que os tempos de vida, que poden ser duracións dunha enfermidade, dun artigo de consumo (coches, teléfonos, ordenadores, etc.) ou calquera outro tempo entre dous eventos, normalmente requiren de certo seguimento. Se ese seguimento se interrompe, so coñeceremos que o tempo durou polo menos ata o momento da perda do seguimento. Nestas condicións pode seguir interesando considerar o efecto dalgunha variable sobre o tempo de vida. Por exemplo, pode interesar saber se a idade do ou da doente inflúe no tempo de curación dunha lesión. Este traballo consiste en revisar as técnicas de estimación da regresión lineal cando a variable resposta está censurada. Exporanse os métodos xa existentes, estudaranse as súas propiedades mediante simulacións, e ilustraranse con datos reais.
Recomendacións
Ter un coñecemento básico do programa estatístico R.

Índice xeral

Resumo

Os datos censurados son bastante habituais no contexto da Análise de Supervivencia, que é unha parte da Estatística que se centra en modelizar o tempo que transcorre ata que ocorre un determinado suceso. Un exemplo notable desta situación é o tempo de vida dunha certa enfermidade que se pode definir como o tempo que pasa dende o comezo dun experimento ata que ocorre un determinado suceso de interese que chamaremos morte ou fracaso (falecemento do/da doente, fin do estudo, perda de información sobre o/a doente,...). Polo tanto, o fenómeno da censura xorde cando existe unha limitación na información que temos sobre as variables de interese dun determinado modelo posto que a partir dun certo intre non podemos observalas.

Neste traballo estudaremos as propiedades teóricas dos diferentes métodos que se empregan para estimar os parámetros asociados a un modelo de regresión no caso de de que a variable resposta sexa censurada pola dereita. Empregaremos modelos de regresión lineais simples para intentar explicar a relación dun par de variables e observaremos como non se poden empregar os mesmos métodos que para o caso de datos completos.

Unha vez expostos os diferentes métodos, compararemos estes estimadores mediante un estudo de simulación empregando o método de Monte Carlo para comprobar que método nos proporciona mellores resultados. Para medir a calidade dos diferentes estimadores dispoñibles na literatura empregaremos o erro cadrático medio.

Finalmente, para rematar este TFG, realizaremos unha aplicación a datos reais que nos permitirá ilustrar o comportamento na práctica dos diferentes métodos estudados ao longo deste traballo. Tanto o estudo de simulación como a aplicación a datos reais levaranse a cabo empregando o *software* estatístico libre R.

Palabras chave: Datos censurados, modelos de regresión, simulación Monte Carlo.

Resumen

Los datos censurados son bastante habituales en el contexto de la Análisis de Supervivencia, que es una parte da Estadística que se centra en modelar el tiempo que transcurre hasta que ocurre un determinado suceso. Un ejemplo notable de esta situación es el tiempo de vida de una cierta enfermedad que se puede definir como el tiempo que pasa desde el comienzo de un experimento hasta que ocurre un determinado suceso de interés que llamaremos muerte o fracaso (fallecimiento del o de la paciente, fin del estudio, pérdida de la información sobre el/la paciente, ...). Por tanto, el fenómeno de censura aparece cuando existe una limitación en la información que tenemos sobre las variables de interés de un determinado modelo puesto que a partir de un cierto momento no podemos observarlas.

En este trabajo estudiaremos las propiedades teóricas de los diferentes métodos que se utilizan para estimar los parámetros asociados a un modelo de regresión en el caso de que la variable respuesta sea censurada por la derecha. Utilizaremos modelos de regresión lineales simples para intentar explicar la relación de un par de variables y observaremos cómo no se pueden utilizar los mismos métodos que para el caso de datos completos.

Una vez expuestos los diferentes métodos, compararemos estos estimadores mediante un estudio de simulación utilizando el método de Monte Carlo para comprobar qué método nos proporciona mejores resultados. Para medir la calidad de los diferentes estimadores disponibles en la literatura utilizaremos el error cuadrático medio.

Finalmente, para acabar este TFG, realizaremos una aplicación a datos reales que nos permitirá ilustrar el comportamiento en la práctica de los diferentes métodos estudiados a lo largo de este trabajo. Tanto el estudio de simulación como la aplicación a datos reales se llevará a cabo utilizando el *software* estadístico libre  .

Palabras clave: Datos censurados, modelos de regresión, simulación Monte Carlo.

Abstract

Censored data is quite common in the context of Survival Analysis, which is a part of Statistics that focuses on modeling the time that passes until a certain event occurs. A notable example of this situation is the life time of a certain disease that can be defined as

the time that passes from the beginning of an experiment until a certain event of interest occurs that we will call death or failure (death of the patient, end of the study, loss of information about the patient, ...). Therefore, the phenomenon of censorship appears when there is a limitation on the information that we have on the variables of interest of a certain model since, from a certain moment, we can not observe them.

In this work we will study the theoretical properties of the different methods that are used to estimate the parameters associated with a regression model in the case that the response variable is right censored. We will use simple linear regression models to try to explain the relationship of a pair of variables and we will observe how the same methods can not be used as for the case of complete data.

Once the different methods are presented, we will compare these estimators through a simulation study using the Monte Carlo method to check which method gives us better results. To measure the quality of the different estimators available in the literature, we will use the mean square error.

Finally, we will perform a real data application that will allow us to illustrate the behavior in practice of the different methods studied throughout this work. Both the simulation study and the real data application to will be carried out using the statistical software  .

Key words: Censored data, regression models, simulation Monte Carlo.

Capítulo 1

Introdución

Unha **variable aleatoria** Y representa unha certa característica dunha poboación que varía entre os diferentes individuos da mesma e queda determinada pola súa **función de distribución**, que se define da seguinte maneira:

$$F(y) = \mathbb{P}\left(Y \leqslant y\right).$$

Pódese dar o caso de que haxa outras variables que inflúan nesa variable Y da que partimos. Un exemplo sería considerar as variables altura, peso e talla dunha mesma persoa, onde observamos facilmente que as dúas primeiras variables inflúen na última. Desta forma xorden os modelos de regresión, para observar e intentar explicar esta relación entre variables aleatorias.

Comezaremos repasando os **modelos de regresión** lineal simple vistos na materia de Inferencia Estatística, o que nos servirá de introdución teórica para este traballo. Estudaremos como construír un modelo de regresión para representar a dependencia lineal dunha variable resposta (ou variable dependente), Y, respecto a outra variable explicativa (ou variable independente) X. Un dos principais obxectivos dos modelos de regresión é facer predicións do valor de Y cando se coñece un certo valor de X. Partimos de que un modelo de regresión lineal simple podémolo expresar da seguinte forma:

$$Y = \beta_0 + \beta_1 X + \varepsilon, \tag{1.1}$$

onde ε é o que lle chamaremos erro e verifica que $\mathbb{E}\left(\varepsilon | X = x\right) = 0$. Durante este capítulo estudaremos estimadores dos parámetros β_0 e β_1, que se corresponden coa ordenada no orixe e coa pendente da recta de regresión, respectivamente.

1.1. Hipóteses do modelo de regresión lineal simple

Para poder estimar o modelo (1.1) necesitamos as seguintes hipóteses:

- **Linealidade.** Debido a que a función de regresión é unha liña recta podemos expresar este modelo como aparece na ecuación (1.1).
- **Homocedasticidade.** Para cada valor da variable explicativa x, a varianza do erro debe ser constante, é dicir:

$$\text{Var}(\varepsilon|X = x) = \sigma^2 \text{ para todo } x.$$

- **Independencia.** Os erros teñen que ser independentes entre si.
- **Normalidade.** O erro ten distribución normal de media cero e varianza σ^2, é dicir:

$$\varepsilon \in N\left(0, \sigma^2\right)^1.$$

Para poder estimar β_0 e β_1 necesitamos unha mostra de datos que se obteñen ou ben dun deseño fixo ou dun deseño aleatorio. No deseño fixo fíxanse os valores da variable explicativa de forma que temos unha mostra do tipo $\{(x_1, Y_1), ...\, (x_n, Y_n)\}$. En canto ao deseño aleatorio, tanto a variable explicativa como a variable resposta son aleatorias, o que nos da unha mostra do tipo $\{(X_1, Y_1), ...\, (X_n, Y_n)\}$. Para poder facer inferencia sobre os parámetros desexados usaremos un deseño de tipo fixo. Neste traballo consideraremos que $\{W_1, ..., W_n\}$ son os datos e $W^{(1)}, ..., W^{(p)}$ son as correspondentes variables aleatorias.

Resumindo, traballaremos cun modelo de regresión lineal simple, homocedástico, con errores normais e independentes e baixo deseño fixo.

1.2. Estimación dos parámetros

Para estimar os parámetros β_0, β_1 e σ^2 asociados a un modelo de regresión lineal simple supoñeremos as hipóteses antes explicadas. Para a predición do valor da variable resposta Y a partir do valor da variable explicativa X, temos os seguintes erros, chamados **residuos** da regresión:

$$\widehat{\varepsilon}_i = Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i \qquad \text{sendo } i \in \{1, ..., n\}.$$

[1] Cando escribimos $N\left(\mu, \sigma^2\right)$ referímonos a unha normal de media μ e varianza σ^2, sendo a normal unha das distribucións de probabilidade máis frecuentes en Estatística.

Para efectuar a estimación faremos uso do **método de mínimos cadrados** cuxo obxectivo é minimizar a suma de residuos ao cadrado. Así, temos que buscar $\widehat{\beta}_0$ e $\widehat{\beta}_1$ de forma que fagan mínima esta suma, que vén dada por:

$$\sum_{i=1}^{n}\left(Y_i-\widehat{\beta}_0-\widehat{\beta}_1x_i\right)^2=\min_{\beta_0,\beta_1}\sum_{i=1}^{n}\left(Y_i-\beta_0-\beta_1x_i\right)^2.$$

Se derivamos a expresión anterior con respecto a β_0 e β_1 e logo igualamos a cero, obtemos as seguintes expresións:

$$\boxed{\widehat{\beta}_0=\overline{Y}-\frac{S_{xY}}{S_x^2}\overline{x}} \qquad \boxed{\widehat{\beta}_1=\frac{S_{xY}}{S_x^2}}$$

Por último estimaremos a varianza do erro σ^2 do seguinte xeito:

$$\widehat{\sigma}^2=\frac{1}{n-2}\sum_{i=1}^{n}\widehat{\varepsilon}_i^2=\frac{1}{n-2}\sum_{i=1}^{n}\left(Y_i-\widehat{\beta}_0-\widehat{\beta}_1x_i\right)^2$$

1.3. Propiedades dos estimadores

Unha vez coñecidas as expresións dos estimadores, vainos ser interesante estudar as súas propiedades.

Propiedades de $\widehat{\beta_1}$

Para calcular a esperanza de forma máis sinxela, expresaremos $\widehat{\beta}_1$ da seguinte forma:

$$\widehat{\beta}_1=\frac{S_{xY}}{S_x^2}=\frac{\sum_{i=1}^{n}\left(x_i-\overline{x}\right)\left(Y_i-\overline{Y}\right)}{nS_x^2}=\sum_{i=1}^{n}\frac{\left(x_i-\overline{x}\right)}{nS_x^2}(Y_i-\overline{Y})=\sum_{i=1}^{n}\omega_i\left(Y_i-\overline{Y}\right)$$

sendo $\omega_i=\dfrac{(x_i-\overline{x})}{nS_x^2}$ os pesos que so dependen da variable explicativa e, como estamos supoñendo que traballamos cun deseño fixo, estes pesos non son aleatorios. Así, usando propiedades da media, podemos facer:

$$\mathbb{E}\left(\widehat{\beta}_1\right)=\mathbb{E}\left(\sum_{i=1}^{n}\omega_i\left(Y_i-\overline{Y}_i\right)\right)=\sum_{i=1}^{n}\omega_i\mathbb{E}\left(Y_i-\overline{Y}\right)=\sum_{i=1}^{n}\underbrace{\frac{(x_i-\overline{x})}{nS_x^2}}_{\omega_i}\underbrace{\beta_1\left(x_i-\overline{x}\right)}_{\mathbb{E}(Y_i-\overline{Y})}=\beta_1$$

Para demostrar as últimas igualdades temos que ter en conta que:

- $\mathbb{E}\left(Y_i\right)=\beta_0+\beta_1x_i$.

- $\mathbb{E}\left(\overline{Y}\right) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left(Y_i\right) = \frac{1}{n}\sum_{i=1}^{n}\left(\beta_0+\beta_1 x_i\right) = \beta_0+\beta_1\overline{x}.$
- En consecuencia temos $\mathbb{E}\left(Y_i-\overline{Y}\right) = \beta_0+\beta_1 x_i-\beta_0-\beta_1\overline{x} = \beta_1\left(x_i-\overline{x}\right).$
- Na última igualdade úsase que $S_x^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2$.

Para poder calcular a varianza, imos expresar $\widehat{\beta}_1$ da seguinte forma:

$$\widehat{\beta}_1 = \sum_{i=1}^{n}\omega_i\left(Y_i-\overline{Y}\right) = \sum_{i=1}^{n}\omega_i Y_i.$$

debido a que $\sum_{i=1}^{n}\omega_i = 0$. Entón,

$$\mathrm{Var}\left(\widehat{\beta}_1\right) = \mathrm{Var}\left(\sum_{i=1}^{n}\omega_i Y_i\right) = \sum_{i=1}^{n}\omega_i^2\ \mathrm{Var}(Y_i) = \sum_{i=1}^{n}\frac{\left(x_i-\bar{x}\right)^2}{n^2 S_x^4}\sigma^2 = \frac{\sigma^2}{nS_x^2}.$$

Para explicar estas igualdades recordamos que estamos traballando coa hipótese de independencia dos erros e homoceasticidade e usamos propiedades básicas do operador varianza.

Finalmente, como β_1 é combinación lineal de $Y_1,\ldots,Y_n$ que son variables independentes e normais, este estimador ten distribución normal. En resumo:

$$\boxed{\widehat{\beta}_1 \in N\left(\beta_1, \frac{\sigma^2}{nS_x^2}\right)}$$

Propiedades de $\widehat{\beta}_0$

Calcularemos a media e a varianza de forma análoga ao caso anterior. Debido a que $\widehat{\beta}_0 = \overline{Y}-\widehat{\beta}_1\overline{x}$, temos que:

$$\mathbb{E}\left(\widehat{\beta}_0\right) = \mathbb{E}\left(\overline{Y}\right)-\overline{x}\mathbb{E}\left(\widehat{\beta}_1\right) = \beta_0+\beta_1\overline{x}-\overline{x}\beta_1 = \beta_0.$$

Para calcular agora a varianza, usaremos que $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$, $\overline{x}\widehat{\beta}_1 = \sum_{i=1}^{n}\overline{x}\omega_i Y_i$ e así temos:

$$\widehat{\beta}_0 = \frac{1}{n}\sum_{i=1}^{n} Y_i - \sum_{i=1}^{n}\overline{x}\omega_i Y_i = \sum_{i=1}^{n}\left(\frac{1}{n}-\overline{x}\omega_i\right)Y_i.$$

Así, tendo en conta as hipóteses básicas do modelo de regresión lineal simple, verifícase que

$$\mathrm{Var}\left(\widehat{\beta}_0\right) = \sum_{i=1}^{n}\left(\frac{1}{n}-\overline{x}\omega_i\right)^2\mathrm{Var}\left(Y_i\right) = \sigma^2\sum_{i=1}^{n}\left(\frac{1}{n^2}+\overline{x}^2\omega_i^2-\frac{2\bar{x}\omega_i}{n}\right) = \sigma^2\left(\frac{1}{n}+\frac{\bar{x}^2}{nS_x^2}\right),$$

onde para a última igualdade usamos que $\sum_{i=1}^{n} \omega_i = 0$ e $\sum_{i=1}^{n} \omega_i^2 = \frac{1}{nS_x^2}$.

Finalmente, debido a que $\widehat{\beta_0}$ é combinación lineal de $Y_1, \ldots, Y_n$, tal como sucedía no caso de $\widehat{\beta_1}$, temos que $\widehat{\beta_0}$ ten unha distribución normal. Polo tanto, podemos concluír:

$$\boxed{\widehat{\beta_0} \in N\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\overline{x}^2}{nS_x^2}\right)\right)}$$

Propiedades de $\widehat{\sigma}^2$

Aínda que neste caso non entraremos en detalles, o estimador da varianza do erro segue unha distribución do tipo chi-cadrado:

$$\boxed{\frac{(n-2)\,\widehat{\sigma}^2}{\sigma^2} \in \chi^2_{n-2}}$$

Nótese que cando estimamos a varianza, dividimos entre $n-2$ en lugar de facelo entre n para que o estimador σ^2 sexa insesgado.

1.4. Regresión lineal múltiple

Unha vez visto o modelo de regresión lineal simple, podemos estendelo a situacións máis complexas onde hai máis dunha variable explicativa, o que se coñece como regresión lineal múltiple. Neste modelo temos unha variable resposta Y e unha colección de variables explicativas $X^{(1)}, \ldots, X^{(p-1)}$. Esta clase de modelos podemos escribilos en forma matricial do seguinte xeito:

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{1,1} & \cdots & x_{1,p-1} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n,1} & \cdots & x_{n,p-1} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_{p-1} \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

A expresión simplificada deste modelo é:

$$Y = \mathbb{X}\beta + \varepsilon,$$

onde Y é o vector das variables respostas, $\mathbb{X}$ é unha matriz de $n \times p$ elementos, $\beta \in \mathbb{R}^p$ é o vector de coeficientes e ε é o vector dos erros que verifica $\varepsilon \in N_n\left(0, \sigma^2 I_n\right)$, sendo I_n a matriz identidade. Para estimar β, aplicaremos o método de mínimos cadrados que ao igual que no caso do modelo lineal simple ten como obxectivo minimizar a suma de residuos ao cadrado, é dicir:

$$\widehat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} \left(Y_i - x_i\beta\right)^2,$$

sendo x_i a fila i-ésima da matriz $\mathbb{X}$. Este problema tamén se pode escribir en notación matricial de maneira equivalente como:

$$\widehat{\beta} = \arg\min_{\beta} (Y - \mathbb{X}\beta)' (Y - \mathbb{X}\beta) .$$

Se derivamos con respecto β e igualamos a cero, obtemos:

$$\mathbb{X}'\mathbb{X}\beta = \mathbb{X}'Y,$$

e a súa solución é o estimador de β, que ven dado por:

$$\boxed{\widehat{\beta} = (\mathbb{X}'\mathbb{X})^{-1} \mathbb{X}'Y.}$$

Non escribiremos a demostración detallada pero é importante destacar as seguintes propiedades deste estimador:

- Media: $\mathbb{E}\left(\widehat{\beta}\right) = \beta$, é dicir, trátase dun estimador inesgado.
- Covarianza: Cov $\left(\widehat{\beta}, \widehat{\beta}\right) = \sigma^2 (\mathbb{X}'\mathbb{X})^{-1}$.
- Distribución límite:

$$\boxed{\widehat{\beta} \in N_p \left(\beta, \sigma^2 (\mathbb{X}'\mathbb{X})^{-1}\right)}$$

Capítulo 2

Datos censurados

Neste capítulo explicaremos que son os **datos censurados** moi empregados no ámbito da **Análise de Supervivencia**. Comezaremos dando unha idea sobre que é a censura e os datos censurados grazas a un exemplo motivador para logo explicar estes termos con máis profundidade.

No ámbito da Bioestatística[1] podemos atopar moitos exemplos de datos censurados. Imaxinemos que traballamos cun estudo sobre o cancro. Neste caso os/as doentes poden morrer ou abandonar o estudo por diferentes causas e se isto ocorre non imos ter coñecemento dos seus datos completos. Isto é o que se coñece como datos censurados. Empregaremos como exemplo un ensaio clínico, que é un procedemento experimental dun medicamento ou tratamento en persoas para avaliar a súa seguridade ou a súa eficacia. Vexamos un exemplo concreto dun ensaio clínico que podemos atopar en [12]. Na Figura 2.1 representamos o tempo de vida de seis doentes e observamos que entran ao estudo durante un período de 2.5 anos que vai dende comezos do ano 2000 ata mediados do 2002. Os/As individuos/as son seguidos durante 4.5 anos ata finais do 2007, cando remata o ensaio clínico.

Na Figura 2.1, as rectas verticais representan o comezo do ensaio, o peche do período para poder entrar ao estudo e o seu remate, respectivamente. Cada liña horizontal representa a un/unha individuo/a onde os puntos negros marcan a súa entrada ao estudo, os círculos representan os eventos censurados e as aspas denotan a morte do/da doente.

Neste exemplo observamos un caso de censura xa que no momento que remata o estudo tres doentes (D1, D3 e D4) seguían vivos/as. Polo tanto, para estes/as tres doentes

[1] A Bioestatística é unha rama da Estadística aplicada ás Ciencias da Vida como son a Bioloxía e a Medicina.

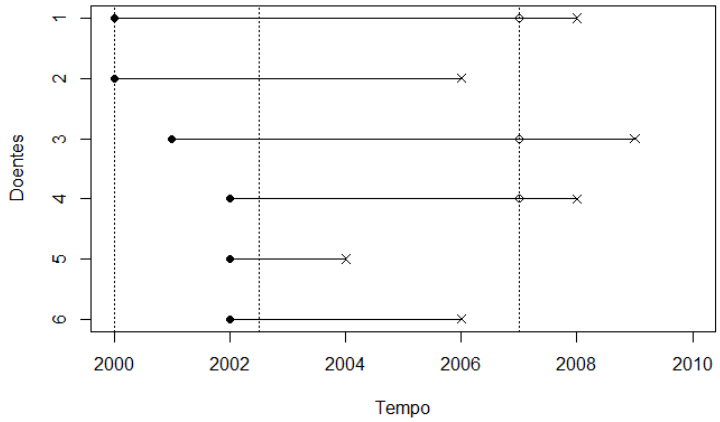


Figura 2.1: Ensaio clínico sobre seis doentes onde as rectas verticais representan o comezo do ensaio, o peche do período para poder entrar ao estudo e o seu remate, respectivamente. Cada liña horizontal representa a un/unha individuo/a onde os puntos negros marcan a súa entrada ao estudo, os círculos representan os eventos censurados e as aspas denotan a morte do/da doente.

non temos información completa sobre o seu tempo de vida e polo tanto estes datos serán censurados. No caso do/da doente D1, por exemplo, sabemos que sobreviviu polo menos 7 anos pero non sabemos canto tempo máis vai vivir. Na Figura 2.1 temos os datos da súa morte (representada cunha aspa) pero este valor non sería coñecido no momento do estudo. En resumo, temos información completa de tres doentes dende que comeza o ensaio ata que morren e temos outros/as tres doentes que serán censurados.

Nun estudo nestas circunstancias, no que non se coñecen todos os datos, poderíamos pensar en traballar so cos datos que temos e obviar o resto. Imos ver graficamente que ocorrería nesta situación empregando a función de distribución empírica. Dada unha mostra $\{Y_1, \ldots, Y_n\}$, recordemos que a función de distribución empírica dunha variable aleatoria Y é da seguinte forma:

$$\widehat{F}_n(y) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(Y_i \leq y). \tag{2.1}$$

Na Figura 2.2 observamos dúas gráficas. A parte (a) representa a función de distribución dunha variable normal $N(0,1)$ para unha mostra de tamaño $n = 100$, mentres que a parte (b) é para tamaño $n = 1000$. O trazo en azul correspóndese coa función de distribución empírica para os datos cunha censura para valores menores que 1 (quedamos cos datos que cumpran $y \leq 1$) mentres que o vermello representa a función de distribución

empírica dos datos completos (é dicir, sen presenza de censura). A curva negra representa a función de distribución teórica dunha $N(0,1)$ e observamos como a curva vermella aproxima ben ao modelo teórico mentres que a azul non. Desta forma podemos ver que no caso dos datos censurados non podemos quedarnos cos datos que temos ata un certo tempo t xa que eses datos non se corresponde coa realidade. O problema tampouco é do tamaño da mostra xa que observamos a mesma situación no gráfico da parte (a) e no da parte (b). Polo tanto, xorde a necesidade de buscar outras opcións para traballar con estes datos. Por exemplo, en lugar de empregar a función de distribución empírica, no caso de datos censurados usaremos a función Kaplan-Meier da que falaremos neste mesmo capítulo na Sección 2.5.

Agora definiremos a censura e os datos censurados formalmente.

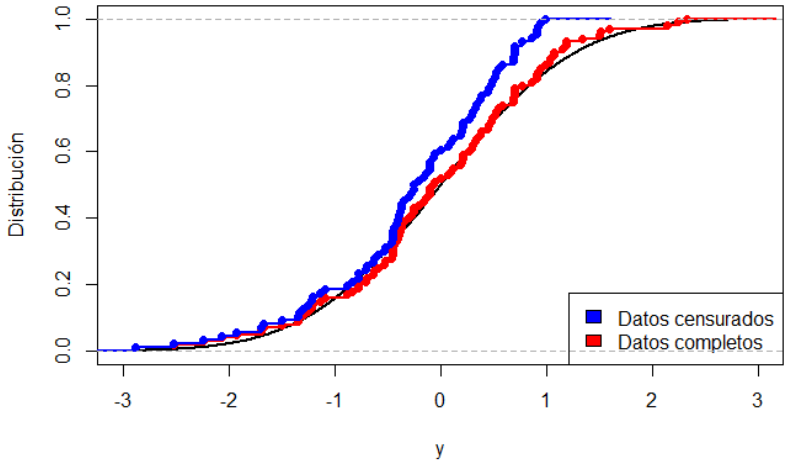
Definición 2.1. Denomínase **censura** ao fenómeno que afecta ás variables de interese nun estudo cando existe unha limitación na información que temos delas. Os **datos censurados** son observacións que non poden ser cuantificadas, dado que so se coñece que o seu valor se atopa por debaixo ou por arriba dunha cota determinada ou ben que está incluído nun intervalo.

2.1. Introdución á Análise de Supervivencia

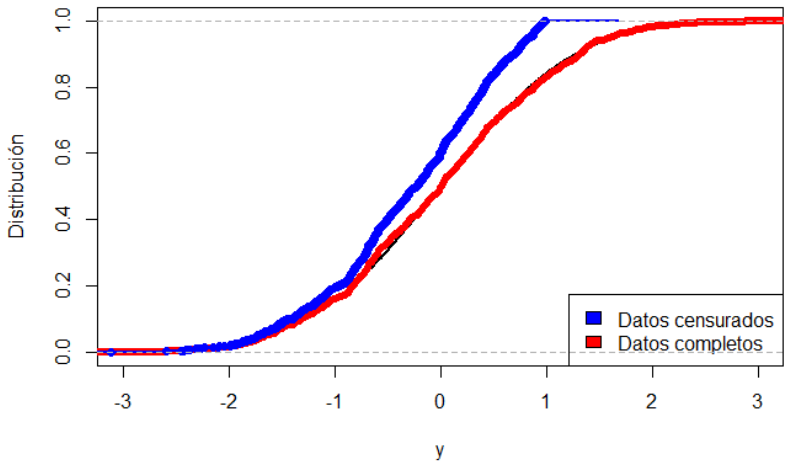
Supoñamos que estamos interesados en estudar unha determinada variable Y que representa o tempo que pasa dende o comezo do experimento ata que ocorre un determinado suceso de interese que chamaremos morte ou fracaso. Así, cando nos refiramos á variable no tempo t, estamos falando da variable Y. Este suceso tamén se pode entender como algo positivo, como por exemplo o tempo que pasa dende que un/unha doente entra nun ensaio clínico ata que responde favorablemente a un tratamento. O conxunto de técnicas estatísticas que se empregan para analizar este tipo de datos coñécense como a **Análise de Supervivencia.** Debido á presenza de censura, a Análise de Supervivencia tamén se coñece como análise de datos censurados.

A principal característica da Análise de Supervivencia é que a variable Y é unha variable discreta ou continua non negativa e representa o tempo dende un inicio ata unha fin definidos. Outra característica importante xurde cando o comezo ou a fin dun evento non se observan completamente. Estes fenómenos coñécense como censura pola dereita e censura pola esquerda.

- **Censura pola dereita:** aparece cando o extremo final é so coñecido por exceder un



(a) Tamaño: $n = 100$



(b) Tamaño: $n = 1000$

Figura 2.2: Representación gráfica da estimación da función de distribución dunha variable normal no caso de datos censurados a partir de $y = 1$ e no caso de datos completos.

valor particular. Formalmente, sexa Y unha variable que representa o tempo ata o fracaso e C unha variable que representa o tempo para un evento censurado. Entón definimos $Z = \text{mín}\,(Y, C)$ e $\delta = \mathbb{I}[Y \leqslant C]$, que vai ser o indicador da censura. Desta forma, $\delta = 0$ se Z é un tempo censurado e $\delta = 1$ se Z é o tempo non censurado que se observa completamente. A Figura 2.1 mostra un exemplo de censura pola dereita.

- **Censura pola esquerda:** é menos frecuente e dáse cando os eventos teñen lugar antes do punto de inicio do estudo. Desta forma, o tempo de censura será o tempo de inicio do período de seguimento e non se coñece con exactitude.

Ao longo deste traballo empregaremos datos censurados pola dereita, polo cal asumiremos que $Z = \text{mín}\,(Y, C)$ denota a variable de interese observada e $\delta = \mathbb{I}[Y \leqslant C]$ será a indicadora da censura.

O obxectivo da Análise de Supervivencia é estimar a función de distribución (que veremos na Sección 2.3), comparar dúas ou máis distribucións de supervivencia e avaliar os efectos de certos factores sobre a variable Y. As técnicas que estudaremos teñen importantes similitudes coa clásica regresión lineal en media, coa importante diferenza de que a variable resposta é unha variable censurada.

2.2. Tipos de censura

A partir de agora, como xa dixemos antes, cando falemos de censura estarémonos referindo a censura pola dereita. Esta censura divídese nos seguintes tipos, que explicaremos a partir de exemplos:

Tipo I: Censura por tempo

O tempo de censura está pre-definido. Nun estudo para deixar de fumar, séguese a cada doente dende que comeza o ensaio ata que sofre unha recaída, volve fumar, ou ben se despois de 180 días non se produce unha recaída. Os/As doentes que aos 180 días seguen sen fumar son censurados. Este exemplo podémolo ver en [12].

Tipo II: Censura por número de fallos

Este caso ocorre cando os obxectos experimentais son seguidos ata que unha proporción deles fracasan. Dáse nas áreas da Biomedicina ou no ámbito da industria, onde o tempo de fallo dun dispositivo é o que máis interesa. Un exemplo deste tipo de censura pode darse cando no ámbito da industria realizamos un estudo que remata

despois de que 25 de 100 dispositivos fallen. Neste caso os restantes serían censurados, polo que so unha porcentaxe dos dispositivos son observados.

Tipo III: Censura aleatoria

Recibe este nome debido a que o tempo de censura está determinado por un fenómeno aleatorio, é dicir, o/a investigador/a non ten ningún tipo de control sobre a aparición da censura. No caso da Biomedicina unha causa aleatoria de censura é debido a que o/a doente abandone o ensaio clínico, coa correspondente perda do seu seguimento, ou a morte por algún motivo que non estea relacionada co evento de interese.

2.3. Funcións que caracterizan unha variable censurada

É ben sabido que unha variable aleatoria continua queda caracterizada cando se coñece a súa función de densidade ou ben a súa función de distribución. Non obstante, na Análise de Supervivencia é usual considerar outras funcións que serven para caracterizar con outra visión a mesma variable Y. As funcións máis importantes son a función de Supervivencia e a función de risco[2].

2.3.1. Función de Supervivencia

A función de supervivencia defínese como a probabilidade de que un/unha individuo/a sobreviva durante un tempo igual ou maior que t. Formalmente podemos escribir:

$$S(t) = \mathbb{P}(Y > t) = 1 - F(t) \quad \text{con} \quad 0 < t < \infty, \tag{2.2}$$

onde recordemos que F representa a función de distribución que ven dada por:

$$F(t) = \mathbb{P}(Y \leqslant t),$$

A función de supervivencia ten as seguintes características:

- A tempo inicial 0 toma o valor 1.
- Decrece ou permanece constante ao longo do tempo.
- Nunca toma valores por debaixo de 0.
- A función é continua pola dereita.

A función de supervivencia normalmente aparece definida en termos da función de risco, que definiremos a continuación.

[2] A función de risco coñécese habitualmente tamén como función de *Hazard*.

2.3.2. Función de risco

A función de risco podémola definir como a probabilidade de que, dado un/unha individuo/a que sabemos que sobreviviu ata un tempo t, o/a individuo/a morra no seguinte intervalo de tempo, dividido pola lonxitude deste intervalo. Podémolo expresar formalmente da seguinte forma:

$$h(t) = \lim_{\delta \to 0} \frac{\mathbb{P}(t < Y < t + \delta | Y > t)}{\delta}. \tag{2.3}$$

A función de risco está relacionada coa función de densidade e a función de supervivencia mediante a seguinte expresión:

$$h(t) = \frac{f(t)}{S(t)},$$

onde

$$f(t) = \frac{\partial}{\partial t} F(t) = -\frac{\partial}{\partial t} S(t). \tag{2.4}$$

A función acumulada de risco podémola definir como a área baixo a función de risco ata un instante de tempo t, de maneira análoga a como podemos definir a función de distribución en función da súa función de densidade, é dicir:

$$H(t) = \int_0^t h(u)\ du = -\log S(t).$$

Desta forma, a función de supervivencia podémola definir en termos da función de risco como:

$$S(t) = exp\left(-\int_0^t h(u)\, du\right) = exp\left(-H(t)\right).$$

2.4. Medidas características dunha variable censurada

No contexto dos datos censurados, ao igual que ocorre con datos completos, podemos calcular medidas características como son a media e a mediana tal e como detallaremos a continuación.

A **media** de supervivencia é o valor que se espera para o tempo de supervivencia, é dicir:

$$\mu = \mathbb{E}(Y) = \int_0^\infty t f(t)\, dt.$$

Tendo en conta a ecuación (2.4), e usando integración por partes da seguinte forma:

- $u = t \to du = dt$,

- $dv = f(t)dt \rightarrow v = \int f(t)\,dt = \int -\frac{\partial}{\partial t} S(t)dt = -S(t),$

entón podemos reescribir a media como:

$$\mu = \int_0^\infty S\left(t\right)\, dt.$$

A media da supervivencia so está definida se $S(\infty) = 0$, é dicir, cando todas as persoas involucradas no estudo morren finalmente. Este non sería o caso, por exemplo, cando temos que estudar o tempo para que observemos unha recaída nunha enfermidade como o cancro ou que unha parte dos/as individuos/as, por exemplo c, se curen. Neste caso, teríamos que $S(\infty) = c$ e a área baixo a curva de supervivencia sería infinita.

A **mediana** do tempo de supervivencia defínese como o tempo que ten que pasar para que $S\left(t\right) = \frac{1}{2}$. Se a curva de supervivencia non é continua en $\frac{1}{2}$, entón a mediana será o t máis pequeno tal que $S\left(t\right) \leqslant \frac{1}{2}$. Se a curva de supervivencia nunca baixa de $\frac{1}{2}$ durante o período de observación, entón a mediana neste caso non estaría definida.

2.5. Estimador de Kaplan-Meier

No caso de datos completos temos a coñecida función de distribución empírica vista en (2.1),

$$\widehat{F}_n\left(y\right) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\left(Y_i \leq y\right), \tag{2.5}$$

que nos permite estimar a función de distribución da variable de interese Y.

No caso dos datos censurados, como ilustramos na Figura 2.2, non se pode empregar a función de distribución empírica (2.5) e xorde así a aparición do estimador de Kaplan-Meier que foi proposto en 1958 e podemos ver con máis detalle en [9]. A función de distribución Kaplan- Meier é da seguinte forma

$$\widehat{F}_{KM}\left(t\right) = \sum_{i=1}^{n} W_i \mathbb{I}\left(Z_{(i)} \leq t\right). \tag{2.6}$$

Recordemos que con datos censurados, $\delta_i = \mathbb{I}\left(Y_i \leq C_i\right)$ é a indicadora da censura, sendo $\{Y_1, ..., Y_n\}$ os datos teóricos, C_i o tempo potencial de censura e Z_i os valores observados da variable de interese, $Z_i = \text{mín}\left(Y_i, C_i\right)$. É importante ter en conta que hai que ordenar os datos de forma que $Z_{(1)} \leq Z_{(2)} \leq \cdots \leq Z_{(n)}$, e por iso en (2.6) escribimos $Z_{(i)}$, que é o elemento que ocupa a i-ésima posición. Tamén temos que definir os pesos W_i, coñecidos

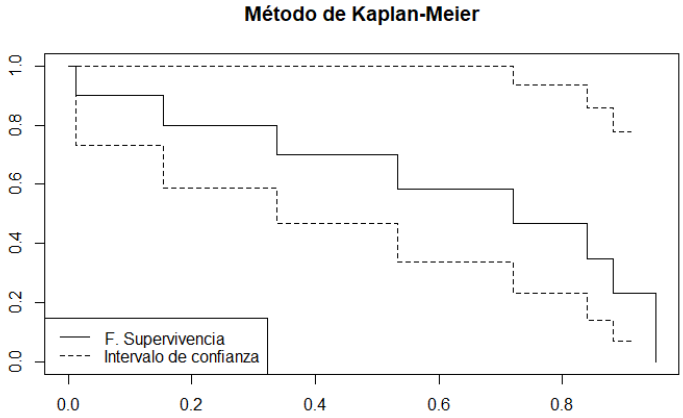


Figura 2.3: Gráfica da aproximación da función de Supervivencia mediante o método de Kaplan-Meier.

como **pesos Kaplan-Meier**:

$$W_i = \frac{\delta_{(i)}}{n-i+1} \prod_{j=1}^{i-1} \left[1 - \frac{\delta_{(j)}}{n-j+1}\right].$$

Se queremos estimar a función de Supervivencia, escribimos

$$\widehat{S}(t) = 1 - \widehat{F}_{KM}(t) = \prod_{t_i \leq t} \left(\frac{n-i}{n-i+1}\right)^{\delta_i} \tag{2.7}$$

Se comparamos (2.6) con (2.5) podemos ver que para a distribución de Kaplan-Meier estamos considerando uns pesos W_i para cada observación en vez de darlle un peso de $\frac{1}{n}$ a todos os datos. Os pesos de Kaplan-Meier para o caso de datos completos toman o valor $\frac{1}{n}$ tal como podemos comprobar mirando a definición dos W_i.

O estimador de Kaplan-Meier está ben definido para tódolos valores de t menores que o maior tempo observado. Porén, se o maior valor observado corresponde a un tempo de vida non censurado, entón a curva de supervivencia para valores de t posteriores é 1. En cambio, se a observación $Z_{(n)}$ é censurada, o valor de $F(t)$ para tempos posteriores non alcanzará o valor 1, que é unha característica propia da función de distribución. Esta situación reflexa un problema de consistencia deste estimador.

Na liña de mellorar o estimador proposto por Kaplan e Meier, destacamos o estimador presuavizado de Kaplan-Meier cuxa principal achega é outorgar diferentes pesos tanto aos

datos censurados como aos non censurados. Neste caso, empregariamos os seguintes pesos

$$W_i^P = \frac{\mathbb{P}(\delta = 1|Z = Z_{(i)})}{n-i+1} \prod_{j=1}^{i-1} \left[1 - \frac{\mathbb{P}(\delta = 1|Z = Z_{(j)})}{n-j+1}\right].$$

Existen diferentes propostas para estimar $\mathbb{P}(\delta = 1|Z)$ como métodos non paramétricos introducidos por [2] ou modelos de regresión loxística introducidos por [3]. Diferentes estudos de simulación amosan que o estimador de Kaplan-Meier con pesos presuavizados W_i^P proporciona mellores resultados que o clásico estimador de Kaplan-Meier, especialmente para tamaños de mostra pequenos.

Capítulo 3

Regresión censurada

Ao longo deste capítulo estudaremos como estimar os parámetros asociados a un modelo de regresión lineal simple cando a variable resposta está censurada pola dereita. É dicir, trataremos que modelas a relación entre unha variable resposta (que denotaremos por Y) e unha variable explicativa (que denotaremos por X) a través dunha recta de regresión da forma:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

onde ε representa o erro do modelo. Supoñemos coñecida unha mostra $\{x_1, \dots, x_n\}$ da variable explicativa baixo deseño fixo. Ademais, dado que a variable resposta será censurada pola dereita, denotaremos $\{Z_1, \dots, Z_n\}$ a mostra da variable resposta observada onde $Z_i = \text{mín}\,(Y_i, C_i)$, sendo C a variable de censura e $\delta = \mathbb{I}[Y \leqslant C]$ é o indicador da censura.

Neste contexto, na literatura podemos atopar diferentes propostas que detallaremos ao longo deste capítulo que describiremos ao longo deste capítulo.

3.1. O estimador de mínimos cadrados

Se estamos na situación de coñecer os datos completos dunha mostra $\{Y_1, \dots, Y_n\}$, podemos estimar os parámetros dun modelo de regresión empregando a técnica de mínimos cadrados tal como se fixo no Capítulo 1 deste traballo. Consideramos o modelo de regresión lineal

$$Y_i = \beta_0^{dc} + \beta_1^{dc} x_i + \varepsilon_i,$$

sendo ε_i os erros para cada $i \in \{1, \ldots, n\}$. Desta forma, estimamos $\beta^{dc} = \left(\beta_0^{dc}, \beta_1^{dc}\right)$[1] do seguinte xeito:

$$\widehat{\beta}^{dc} = \arg\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} \widehat{\varepsilon}_i^{\,2} = \arg\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \beta_0^{dc} - \beta_1^{dc} x_i\right)^2. \tag{3.1}$$

O noso obxectivo é ser capaces de estimar o parámetro β no caso de que a variable resposta sexa censurada de forma similar ao que xa fixemos con datos completos. No contexto de datos censurados, como vimos no capítulo anterior, para estimar a función de distribución (en lugar da función de distribución empírica), temos o estimador Kaplan-Meier

$$\widehat{F}_{KM}(y) = \sum_{i=1}^{n} W_i \, \mathbb{I}(Z_i \leq y). \tag{3.2}$$

onde Z_i denota a variable resposta observada (posto que xa non coñecemos os valores de Y_i) e δ_i permítenos identificar a presenza de datos censurados.

Tendo en mente o estimador de Kapla-Meier xunto co estimador (3.1), podemos pensar en estimar o parámetro $\widehat{\beta}$ asociado a un modelo de regresión con variable resposta Y censurada como:

$$\widehat{\beta} = \arg\min \sum_{i=1}^{n} W_i \, \widehat{\varepsilon}_i^{\,2}, \tag{3.3}$$

sendo $\widehat{\varepsilon}_i = Z_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i$ os residuos do modelo de regresión. Esta técnica para atopar o estimador usando os pesos de Kaplan-Meier foi proposta por Stute en [16] (e por iso no seguinte capítulo lle chamaremos método de Stute) e estudada por Sánchez-Sellero na súa tese, que podemos ver en [14]. En concreto, demostraron que o estimador $\widehat{\beta}$ definido en (3.3) é asintoticamente normal, coa seguinte distribución límite para do modelo de regresión lineal simple que estamos tratando:

$$\sqrt{n}\left(\widehat{\beta} - \beta\right) \xrightarrow{\;d\;} N\left(0, \Omega^{-1} \, \Pi \, \Omega^{-1}\right),$$

considerando a seguinte notación:

$$\Omega = \left\{ \mathbb{E}\left[\frac{\partial m(x)}{\partial \beta_r}, \frac{\partial m(x)}{\partial \beta_s} \right] \right\}_{r,s \in \{0,1\}},$$

sendo $m(x) = \mathbb{E}(Y \mid X = x) = \beta_0 + \beta_1 x, \; x \in \mathbb{R}$, e

$$\Pi = \left(\mathrm{Cov}\left(\eta^{\varphi_r}, \eta^{\varphi_s}\right)\right)_{r,s \in \{0,1\}},$$

[1] O superíndice dc fai referencia a que estamos traballando con datos completos.

onde $\varphi_r(x,y) = [y - m(x)] \dfrac{\partial m(\beta)}{\partial \beta_r}$, e

$$\eta_i^{\varphi} = \varphi(x_i, Z_i)\,\gamma_0(Z_i)\,\delta_i + \gamma_1^{\varphi}(Z_i)\,(1-\delta_i) - \gamma_2^{\varphi}(Z_i),$$

$$\gamma_0(y) = \exp\left\{\int_{-\infty}^{y^-} \frac{\widetilde{H}^0(dZ)}{1-H(Z)}\right\},$$

sendo $F(u)$ e $D(u)$ as funcións de distribución das variables Y e C respectivamente, entón temos:

$$\widetilde{H}^0(y) = \mathbb{P}(Z \leqslant y, \delta = 0) = \int_{-\infty}^{y} (1 - F(u))\, D(du),$$

e

$$\widetilde{H}^{11}(x,y) = \mathbb{P}(X \leqslant x, z \leqslant y, \delta = 1),$$

$$H(y) = \mathbb{P}(Z \leqslant y).$$

Ademais, para φ unha función real medible definida en $\mathbb{R}^2$, definimos:

$$\gamma_1^{\varphi}(y) = \frac{1}{1-H(y)} \int \mathbb{I}_{\{y<w\}} \varphi(x,w)\,\gamma_0(w)\,\widetilde{H}^{11}(dx,dw),$$

e

$$\gamma_2^{\varphi}(y) = \int\int \frac{\mathbb{I}_{\{v<y, v<w\}} \varphi(x,w)\,\gamma_0(w)}{[1-H(v)]^2} \widetilde{H}^0(dv)\,\widetilde{H}^{11}(dx,dw).$$

A demostración deste resultado non é obxectivo deste traballo pero podemos atopar todos os detalles explicados en [14].

3.2. O estimador proposto por Miller

Recordemos primeiro brevemente en que consiste o método de máxima verosimilitude que imos empregar para datos completos. Sexa $\{V_1, \ldots, V_n\}$ unha mostra aleatoria simple dunha variable V con función de distribución F_θ ou función de densidade f_θ. O estimador de máxima verosimilitude (en adiante, EMV) é aquel valor que maximiza a masa de probabilidade (ou densidade) da mostra. Se definimos a función de verosimilitude como

$$L(\theta) = L(V_1, \ldots, V_n; \theta) = f_\theta(V_1, \ldots, V_n) = \prod_{i=1}^{n} f_\theta(Y_i). \tag{3.4}$$

Así, para cada mostra particular $\{Y_1, \ldots, Y_n\}$ a estimación de máxima verosimilitude de β é o valor $\widehat{\beta}_{MV}$ que maximiza a verosimilitude definida en (3.4), é dicir:

$$L\left(V_1, \ldots, V_n; \widehat{\theta}_{MV}\right) = \max_{\theta} L(V_1, \ldots, V_n; \theta).$$

Polo tanto, sexa θ o vector dos parámetros, o procedemento que hai que seguir para obter o **EMV** de θ_j, dada unha mostra $\{V_1, \ldots, V_n\}$ é:

1. Escribir a función de verosimilitude $L(\theta) = L(V_1, \ldots, V_n; \theta)$.

2. Escribir o logaritmo da verosimilitude $l(\theta) = \ln L(\theta)$, xa que o máximo non se ve alterado a través desta transformación e así os cálculos resultan máis sinxelos.

3. Obter o θ_j que cumpra
$$\frac{\partial}{\partial \theta_j} l(\theta) = 0,$$
e denotámolo por $\widehat{\theta}_j$.

Unha vez que xa falamos en termos xerais do método de máxima verosimilitude, vexamos agora que ocorre se estamos na situación dos datos censurados. Nesta sección seguiremos o proceso que aparece no libro de Miller que podemos atopar en [11] nas páxinas 11-14, onde aparece explicado para o caso dun modelo de regresión lineal múltiple e neste traballo adaptámolo para o noso contexto dun modelo lineal simple.

Consideramos o par (Z_i, δ_i), recordando que $Z_i = \text{mín}(Y_i, C_i)$, sendo C_i os valores da variable de censura e $\delta = \mathbb{I}[Y \leqslant C]$ é o indicador da censura. Asociado a este par temos a súa correspondente función de densidade:

$$L(Z_i, \delta_i) = \begin{cases} f(Z_i) & se \quad \delta_i = 1, \\ \\ S(Z_i) & se \quad \delta_i = 0, \end{cases}$$

que tamén se pode expresar da seguinte maneira:

$$L(Z_i, \delta_i) = f(Z_i)^{\delta_i} \; S(Z_i)^{1-\delta_i}, \tag{3.5}$$

sendo $f(Z)$ a función de densidade para os datos sen censura e $S(Z)$ a función de supervivencia para os datos censurados. Consideramos agora a función de verosimilitude de toda a mostra

$$L(\beta) = L(Z_1, \ldots, Z_n; \delta_1, \ldots, \delta_n) = \prod_{i=1}^{n} L(Z_i, \delta_i) = \left(\prod_{U} f(Z_i) \right) \left(\prod_{C} S(Z_i) \right), \tag{3.6}$$

denotando por $\prod_U$ e $\prod_C$ o produto sobre os datos sen censura[2] e os datos censurados respectivamente. Nesta igualdade estamos empregando a independencia dos Z_i e usamos (3.5). Ademais, baixo a suposición de que o tempo censurado e o tempo de supervivencia son independentes, cabe destacar que a función de verosimilitude estaría multiplicada por

[2]Empregaremos a notación que vén do inglés: *uncensored* e *censored*.

unha constante que non depende do parámetro β e como o noso obxectivo é maximizar a función, podemos prescindir desta constante.

Recordemos que estamos considerando o vector dos parámetros $\beta = (\beta_0, \beta_1)$. Para atopar o máx$_\beta L(\beta)$ faremos unha transformación mediante o logaritmo e logo derivamos e igualamos a cero para obter o máximo tal e como detallamos no caso de datos completos. Desta forma, usando propiedades da función logaritmo quédanos:

$$0 = \frac{\partial}{\partial \beta_j} \log L(\beta) = \sum_{i=1}^{n} \frac{\partial}{\partial \beta_j} \log L_\beta(Z_i, \delta_i), \text{ con } j \in \{0, 1\}. \tag{3.7}$$

Se substituímos agora o valor de $L(\beta)$ da ecuación (3.6), temos o seguinte:

$$0 = \sum_{U} \frac{\partial}{\partial \beta_j} \log f_\beta(Z_i) + \sum_{C} \frac{\partial}{\partial \beta_j} \log S_\beta(Z_i), \text{ con } j \in \{0, 1\}.$$

Debido á súa complexidade, para poder resolver esta igualdade e obter o valor de $\widehat{\beta}$ teremos que programar un método iterativo coa axuda dun ordenador. Neste traballo, por exemplo, detallaremos o método de **Newton Rapson** que é un algoritmo iterativo que se emprega con frecuencia para atopar aproximacións dos ceros dunha función $G(x)$ necesitando un punto inicial X_0. A partir dunha iteración pasamos a seguinte e así sucesivamente mediante a seguinte ecuación:

$$X_{n+1} = X_n - \frac{G(X_n)}{G'(X_n)}.$$

Para simplificar a notación escribiremos $L_i(\beta) = L_\beta(Z_i, \delta_i)$ con $i = 1, \ldots, n$. Así, podemos reescribir (3.7) da seguinte maneira:

$$0 = \sum_{i=1}^{n} \frac{\partial}{\partial \beta_j} \log L_i(\beta), \ j = 1, \ldots, p,$$

ou ben

$$0 = \frac{\partial}{\partial \beta} \log L(\beta),$$

sendo

$$\frac{\partial}{\partial \beta} \log L(\beta) = \left(\frac{\partial}{\partial \beta_0} \log L(\beta), \frac{\partial}{\partial \beta_1} \log L(\beta) \right),$$

$$\frac{\partial^2}{\partial \beta^2} \log L(\beta) = \begin{pmatrix} \frac{\partial^2}{\partial \beta_0^2} \log L(\beta) & \frac{\partial^2}{\partial \beta_0 \partial \beta_1} \log L(\beta) \\ \\ \frac{\partial^2}{\partial \beta_1 \partial \beta_0} \log L(\beta) & \frac{\partial^2}{\partial \beta_1^2} \log L(\beta) \end{pmatrix}.$$

Supoñamos que a solución inicial é $\widehat{\beta}^0 = \left(\widehat{\beta}_0^0, \widehat{\beta}_1^0\right)$. Empregaremos o desenvolvemento de Taylor[3] para aproximar esa función e poder escribila da seguinte forma:

$$0 = \sum_{i=1}^{n} \frac{\partial}{\partial \beta_j} \log L_i\left(\widehat{\beta}\right) = \sum_{i=1}^{n} \frac{\partial}{\partial \beta_j} \log L_i\left(\widehat{\beta}^0\right) + \sum_{k=0}^{1} \left(\widehat{\beta}_k - \widehat{\beta}_k^0\right) \sum_{i=1}^{n} \frac{\partial^2}{\partial \beta_k \partial \beta_j} \log L_i\left(\widehat{\beta}^0\right) + \dots,$$

con $j = 0, 1$, e tamén o podemos escribir como

$$0 = \frac{\partial}{\partial \beta} \log L\left(\widehat{\beta}\right) = \frac{\partial}{\partial \beta} \log L\left(\widehat{\beta}^0\right) + \left(\widehat{\beta} - \widehat{\beta}^0\right) \frac{\partial^2}{\partial \beta^2} \log L\left(\widehat{\beta}^0\right) + \dots.$$

Usamos agora o método de Newton-Rapson e temos que a solución será

$$\widehat{\beta}^1 = \widehat{\beta}^0 - \frac{\dfrac{\partial}{\partial \beta} \log L\left(\widehat{\beta}^0\right)}{\dfrac{\partial^2}{\partial \beta^2} \log L\left(\widehat{\beta}^0\right)}. \tag{3.8}$$

Este resultado proporciónanos as bases dunha aproximación, de maneira iterativa, para calcular o EMV. Así, dado un valor inicial $\widehat{\beta}^0$, usamos a ecuación (3.8) para obter unha mellor estimación e repetimos este proceso ata xerar unha sucesión de estimadores que converxen ao EMV $\widehat{\beta}$. Consideraremos a seguinte notación:

- $\dfrac{\partial}{\partial \beta} \log L\left(\widehat{\beta}^0\right)$ será o vector derivada en $\widehat{\beta}^0$.
- $-\dfrac{\partial^2}{\partial \beta^2} \log L\left(\widehat{\beta}^0\right)$ será a matriz de información en $\widehat{\beta}^0$ e denotarémola por $i\left(\widehat{\beta}^0\right)$.

Cabe destacar que

$$\mathbb{E}\left(i\left(\beta\right)\right) = -\mathbb{E}\left(\frac{\partial^2}{\partial \beta_k \partial \beta_j} \log L\left(\beta\right)\right) = I\left(\beta\right),$$

sendo $I\left(\beta\right)$ a **matriz de información de Fisher**. Se agora substituímos a expresión do vector derivada e a matriz de información en (3.8) obtemos o seguinte:

$$\widehat{\beta}^1 = \widehat{\beta}^0 + I^{-1}\left(\widehat{\beta}^0\right) \frac{\partial}{\partial \beta} \log L\left(\widehat{\beta}^0\right). \tag{3.9}$$

Debido á dificultade para estimar a función de densidade ou a función de Supervivencia, podemos asumir que a distribución do erro é normal para realizar a estimación de máxima verosimilitude. Debemos destacar tamén que describimos o procedemento para o caso do modelo lineal simple pero de maneira intuitiva poderiamos estendelo ao contexto de múltiples variables explicativas.

[3] Recordemos que o desenvolvemento de Taylor dunha función $g(x)$ consiste en escribir $g(x) = g(x_0) + (x - x_0)\, g'\left(x_0\right) + \dfrac{(x - x_0)^2}{2!} g''\left(x_0\right) + \dots$

3.3. O estimador proposto por Buckley e James

Na Introdución deste traballo, no Capítulo 1, xa empregamos as ecuacións normais para estimar os parámetros no caso do modelo de regresión lineal con datos completos. Imos modificar estas ecuacións para o caso de observacións censuradas pero primeiro recordemos que no caso de datos completos temos que escoller como estimadores de β_0 e β_1 aqueles valores $\widehat{\beta}_0$ e $\widehat{\beta}_1$ que satisfagan:

$$\sum_{i=1}^{n}\left(Y_i-\widehat{\beta}_0-\widehat{\beta}_1 x_i\right)=0, \tag{3.10}$$

$$\sum_{i=1}^{n}\left(x_i-\overline{x}\right)\left(Y_i-\widehat{\beta}_1 x_i\right)=0, \tag{3.11}$$

sendo $\overline{x}=\dfrac{1}{n}\sum_{i=1}^{n}x_i$.

Para o caso de datos censurados, como non podemos observar todos os datos $\{Y_1,\ldots,Y_n\}$, consideraremos

$$Y_i^{*}=Y_i\,\delta_i+\mathbb{E}\left(Y_i|Y_i>C_i\right)\,\left(1-\delta_i\right),\ i=1,\ldots,n,$$

sendo C_i os valores da variable de censura, $\delta_i=\mathbb{I}[Y_i\leqslant C_i]$ o indicador da censura, onde $\delta_i=0$ para os datos censurados e $\delta_i=1$ para os datos non censurados. Entón, $\mathbb{E}\left(Y_i^{*}\right)=\beta_0+\beta_1 x_i$ e polo tanto:

$$\mathbb{E}\left(\sum_{i=1}^{n}\left(x_i-\overline{x}\right)\left(Y_i^{*}-\beta_1 x_i\right)\right)=0.$$

Por analoxía con (3.11), o ideal sería considerar un estimador $\widehat{\beta}_1$ para o cal se cumpra:

$$\sum_{i=1}^{n}\left(x_i-\overline{x}\right)\left(Y_i^{*}-\widehat{\beta}_1 x_i\right)=0.$$

Debido a que $\mathbb{E}\left(Y_i|Y_i>C_i\right)$ é descoñecida, imos considerar unha aproximación consistente empregando a función de Kaplan-Meier $\widehat{F}_{KM}$, onde

$$\widehat{F}_{KM}\left(\varepsilon\right)=1-\prod_{i;\widehat{\varepsilon}_i\leq\widehat{\varepsilon}}\left(\frac{n-i}{n-i+1}\right)^{\delta_i},$$

sendo $\widehat{\varepsilon}_i\left(\widehat{\beta}_0,\widehat{\beta}_1\right)=Z_i-\widehat{\beta}_0-\widehat{\beta}_1 x_i$, $Z_i=\min\left(Y_i,C_i\right)$ e C_i os valores da variable de censura. Imos substituír as observacións censuradas por

$$\overline{Y}_i\left(\widehat{\beta}_1\right)=\widehat{\beta}_1 x_i+\sum_{U}W_{ik}\left(\widehat{\beta}_1\right)\left(Y_k-\widehat{\beta}_1 x_k\right), \tag{3.12}$$

onde neste caso $\sum_U$ é un sumatorio dos datos non censurados sobre k e

$$W_{ik}\left(\widehat{\beta}_1\right) = \begin{cases} \dfrac{v_k\left(\widehat{\beta}_1\right)}{1-\widehat{F}_{KM}\left(C_i-\widehat{\beta}_1 x_i\right)} & se \quad \widehat{\varepsilon}_i\left(0,b\right) < \widehat{\varepsilon}_k\left(0,\widehat{\beta}_1\right), \\ \\ 0 & \text{outro caso.} \end{cases}$$

onde $v_k\left(\widehat{\beta}_1\right)$ é a masa de probabilidade asignada a función de distribución $\widehat{F}_{KM}\left(\varepsilon\right)$ e Temos que escoller entón un estimador $\widehat{\beta}_1$ que satisfaga:

$$\widehat{\beta}_1 = \frac{\sum_U Y_i\left(x_i-\overline{x}\right)+\sum_C \overline{Y}_i\left(\beta\right)\left(x_i-\overline{x}\right)}{\sum_{i=1}^n \left(x_i-\overline{x}\right)^2}. \tag{3.13}$$

Cabe destacar que isto é equivalente a substituír cada punto censurado (C_i, x_i) polos puntos $\{\ \widehat{\beta}_1 x_i + \left(Y_k - \widehat{\beta}_1 x_k\right), x_i\}\ , k \neq i$, darlle o peso $W_{ik}\left(\widehat{\beta}_1\right)$ e finalmente substituílos nas ecuacións normais. Denotamos agora a parte dereita da ecuación (3.13) por $\gamma\left(\widehat{\beta}_1\right)$. Desta forma, en vez de querer minimizar unha función, o noso obxectivo será buscar un b que cumpra $b = \gamma\left(b\right)$, é dicir, queremos resolver a ecuación

$$\gamma\left(b\right) - b = 0, \tag{3.14}$$

da que $\widehat{\beta}_1$ é raíz. Este problema pódese resolver empregando métodos de optimización e así (3.13) reescribímola como

$$\boxed{\widehat{\beta}_1 = \frac{\sum_k^U Y_k\left\{x_k - \overline{x} + \sum_j^C W_{jk}\left(\widehat{\beta}_1\right)\left(x_j-\overline{x}\right)\right\}}{\eta\left(\widehat{\beta}_1\right)}}$$

sendo

$$\eta\left(\widehat{\beta}_1\right) = \sum_{i=1}^n \left(x_i-\overline{x}\right)^2 - \sum_j^C \left(x_j-\overline{x}\right)\left\{x_j - \widetilde{x}_j\left(\widehat{\beta}_1\right)\right\},$$

onde

$$\widetilde{x}_j\left(\widehat{\beta}_1\right) = \sum_k^U W_{jk}\left(\widehat{\beta}_1\right) x_k.$$

Este proceso lévase a cabo no artigo de Buckley e James que podemos ver con máis detalle en [1]. Unha vez que obtemos $\widehat{\beta}_1$, podemos conseguir de forma análoga $\widehat{\beta}_0$ de forma que

$$\boxed{\widehat{\beta}_0 = \frac{\left\{\sum_U Y_i + \sum_C \overline{Y}_i\left(\widehat{\beta}_1\right)\right\}}{n} - \widehat{\beta}_1\overline{x}.}$$

3.4. O estimador proposto por Jin, Lin e Ying

O estimador obtido por Buckely e James é unha raíz da función de estimación que atopamos na ecuación (3.14) que non é nin continua nin monótona e as súas raíces poden non existir. O algoritmo iterativo de Buckley e James presenta algún problema, entre os que destacan que a converxencia do algoritmo non está garantida, é dicir, en ocasións non se atopa a solución ou esta solución oscila entre dous puntos non óptimos. Ademais, aínda que o algoritmo converxa, non está claro que nos leve a un estimador consistente xa que os resultados teóricos foron establecidos baseándose na hipótese de linearidade local. Debido a estes problemas, Jin, Lin e Ying intentaron mellorar este estimador no artigo que podemos ver con detalle en [8].

Para mellorar o método presentado por Buckley e James, será fundamental o papel que xoga o estimador inicial. Se o estimador inicial é consistente, entón para cada paso m, o estimador obtido na m-éstima iteración tamén será consistente. Ademais, se o estimador é asintoticamente normal, entón o estimador obtido na m-éstima iteración tamén o será. Isto que expomos está demostrado por Ritov ou Lai e Ying en [13] e [10], respectivamente. Así, este novo procedemento lévanos a unha clase de estimadores consistentes e asintoticamente normais.

Desta forma, a idea que temos que seguir é a de darlle un "bo" valor inicial ao problema e aplicar o método iterativo. Este tipo de procedementos empréganse frecuentemente no ámbito da Matemática Aplicada como pode ser no método de Newton-Rapson que xa empregamos neste traballo. Jin, Lin e Ying propoñen escoller este valor inicial a partir da función de peso de Gehan, que se pode calcular aplicando técnicas de programación linear.

Consideraremos a mesma notación que empregamos neste Capítulo 3, exceptuando que neste caso consideraremos a transformación logaritmo para a variable resposta. Por simplicidade de notación imos escribir Y_i aínda que estamos considerando $\log Y_i$. Así, traballaremos co seguinte modelo de regresión linear:

$$Y_i = \beta x_i + \varepsilon_i,$$

é dicir, non imos ter en conta o intercepto. Na Sección 3.3 calculamos o estimador da pendente de Buckley e James, que lle chamaremos $\widehat{\beta}_{BJ}$, que é a raíz de $U(\beta, \beta) = 0$, sendo

$$U(\beta, b) = \gamma(b) - \beta,$$

e recordemos que a función γ é a parte dereita da ecuación (3.13), é dicir,

$$\gamma(b) = \frac{\sum_U Y_i (x_i - \overline{x}) + \sum_C \overline{Y}_i(b)(x_i - \overline{x})}{\sum_{i=1}^n (x_i - \overline{x})^2}.$$

É fácil ver que $U(\beta,\beta)$ non é nin continua nin monótona en β e polo tanto é difícil calcular o estimador deste xeito, especialmente cando β é multidimensional. Podemos linealizar a función de estimación primeiramente dando un valor inicial b e logo resolvendo $U(\beta,b)=0$ para β. Esta operación lévanos a realizar $\beta=\gamma(b)$. Continúase este proceso co seguinte algoritmo iterativo:

$$\widehat{\beta}_{(m)}=\gamma\left(\widehat{\beta}_{(m-1)}\right), \quad m\geq 1. \tag{3.15}$$

Un estimador inicial consistente e asintoticamente normal de β_0 pódese conseguir polo método rank-based de Jin, Lin, Wei e Ying, que podemos ver en [7]. Establecemos o estimador inicial $\widehat{\beta}_{(0)}$ como o estimador tipo Gehan, $\widehat{\beta}_G$, descrito en [4] e que podemos calcular minimizando a seguinte función convexa:

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\delta_i\left\{\varepsilon_i(\beta)-\varepsilon_j(\beta)\right\}^{-},$$

onde $a^{-}=\mathbb{I}\{a<0\}|a|$, e recordemos que $\varepsilon_i(\beta)=Z_i-\beta x_i$, $Z_i=\min(Y_i,C_i)$, C_i os valores da variable de censura, $\delta_i=0$ para os datos censurados e $\delta_i=1$ para os datos non censurados. Este problema de minimización é en realidade un problema de programación lineal simple e para ver máis detalles pódese consultar [7].

Para cada m, en [8] próbase que $\widehat{\beta}_{(m)}$ é consistente e asintoticamente normal. Ademais, $\widehat{\beta}_{(m)}$ é unha combinación linear do estimador tipo Gehan $\widehat{\beta}_G$ e do estimador proposto por Buckley e James $\widehat{\beta}_{BJ}$, de forma que

$$\widehat{\beta}_{(m)}=\left(I-D^{-1}A\right)^m\widehat{\beta}_G+\left\{I-\left(I-D^{-1}A\right)^m\right\}\widehat{\beta}_{BJ}+O_p\left(n^{-\frac{1}{2}}\right), \tag{3.16}$$

onde

- I é a matriz identidade.
- $D:=\lim_{n\to\infty}n^{-1}\sum_{i=1}^{n}(x_i-\overline{x})^2$ é a matriz de pendentes da función de estimación de mínimos cadrados para os datos completos (datos non censurados).
- A é a matriz de pendentes da función estimada de Buckley e James que está definida e explicada en [8] pero neste TFG non imos profundizar sobre esta definición xa que non a empregaremos.

Cando a porcentaxe de censura tende a cero, a matriz A aproxímase a D. Así, o primeiro termo da ecuación (3.16) vai ser cero e cada $\widehat{\beta}_{(m)}$ aproxima ao estimador usual de mínimos cadrados. Se o algoritmo iterativo descrito en (3.15) converxe, entón $\widehat{\beta}_{(m)}$ será o estimador

de Buckley e James. Aínda que a secuencia iterativa non converxa, os estimadores seguen sendo consistentes e asintoticamente normais.

Recordemos que unha hipótese do modelo de regresión lineal simple é a normalidade dos erros, polo que os erros seguen unha distribución normal de media cero e varianza σ^2, é dicir, $\varepsilon \in N\left(0, \sigma^2\right)$, sendo a normal unha función non decrecente. Pódese demostrar (ver [8]) que se a función de distribución do erro é non decrecente (como é o noso caso), entón cando $D - A$ é definida positiva isto implica que $\left(I - D^{-1}A\right)^m$ se aproxima a cero ou que $\widehat{\beta}_{(m)}$ se aproxima a $\widehat{\beta}_{BJ}$ (estimador proposto por Buckley e James) cando m tende a ∞ (para tamaños de mostra grandes).

Capítulo 4

Estudo de simulación

Neste capítulo ilustraremos o comportamento na práctica dos métodos presentados ao longo do Capítulo 3 mediante un estudo de simulación. Para realizar isto seguiremos o **método de Monte Carlo** que foi proposto en 1944 e permite resolver problemas matemáticos mediante a simulación de variables aleatorias. O nome deste método débese ao Casino de Monte Carlo xa que a ruleta é un xerador de números aleatorios que permite simular variables aleatorias.

4.1. Introdución

A intención coa que se fai este estudo de simulación é observar as diferenzas entre entre os diferentes estimadores propostos no capítulo anterior. Para levar a cabo esta comparativa empregaremos os termos de sesgo, varianza e erro cadrático medio que definiremos a continuación.

Definición 4.1. O **sesgo** dun estimador $\widehat{\theta}$ é a diferencia entre a súa esperanza matemática e o valor numérico do parámetro que estamos estimando, é dicir,

$$\text{Sesgo}\left(\widehat{\theta}\right) = \mathbb{E}\left(\widehat{\theta}\right) - \theta.$$

Dise que un estimador é insesgado se o seu sesgo é nulo.

Definición 4.2. A **varianza** é unha medida de dispersión que representa a variabilidade dunha variable aleatoria respecto a súa media. Podémola calcular como:

$$\text{Var}\left(\widehat{\theta}\right) = \mathbb{E}\left[\left(\widehat{\theta} - \mathbb{E}\left(\widehat{\theta}\right)\right)^2\right].$$

Definición 4.3. O erro cadrático medio (en adiante, **ECM**) defínese como

$$\text{ECM}\left(\widehat{\theta}\right) = \text{Sesgo}\left(\widehat{\theta}\right)^2 + \text{Var}\left(\widehat{\theta}\right).$$

Para a realización deste estudo de simulación empregaremos o *software* estatístico libre (pódese ver máis información en https://cran.r-project.org/). Ademais, empregamos diferentes librarías de que teñen implementados os diferentes métodos que estudamos ao longo deste traballo. Destacamos as seguintes:

survival

Esta libraría contén todas as técnicas básicas da Análise de Supervivencia. Usamos a función `Surv`(time, status) para crear unha variable tipo Supervivencia sendo, no noso caso, *time* a variable resposta observada Z e *status* a nosa δ (indicadora de censura). Desta forma obtemos unha variable que podemos usar como variable resposta para as fórmulas nalgún método que xa comentaremos. Con `survfit`() podemos estimar curvas de supervivencia empregando o estimador proposto por Kaplan-Meier. Pódese ver máis información sobre esta libraría en

https://cran.r-project.org/web/packages/survival/survival.pdf

condSURV

Permítenos calcular os pesos Kaplan-Meier e os pesos Kaplan-Meier presuavizados que introducimos na Sección 2.5. En concreto usamos a función `KMW`() para os clásicos pesos Kaplan-Meier e a función `PKMW`() para os pesos suavizados. Pódese ver máis información sobre esta libraría en

https://cran.r-project.org/web/packages/condSURV/condSURV.pdf

SMNCensReg

Empregaremos a función `CensReg.SMN`() para estimar os parámetros asociados a un modelo regresión con variable resposta censurada pola dereita usando o método de máxima verosimilitude (desenrolado na Sección 3.2) e asumindo que o erro ten unha distribución normal. Esta función tamén permite asumir outras distribucións para o erro como a T de Student. Pódese ver máis información sobre esta libraría en

https://cran.r-project.org/web/packages/SMNCensReg/SMNCensReg.pdf

rms

Emprégase para estimar os parámetros asociados a un modelo regresión con variable

resposta censurada pola dereita empregando o método de Buckley e James (desenrolado na Sección 3.3). Usaremos a función `bj()` para estimar os parámetros mediante este método. Pódese ver máis información sobre esta libraría en

https://cran.r-project.org/web/packages/rms/rms.pdf

lss2

Usamos a función `lss()` para estimar os parámetros asociados a un modelo regresión con variable resposta censurada pola dereita empregando o método proposto por Jin, Lin e Ying (desenrolado na Sección 3.4). Debemos destacar que este método so permite considerar modelo de regresión sen intercepto. Este método de estimación foi recentemente engadido en R, a finais do 2019. Pódese ver máis información sobre esta libraría en

https://cran.r-project.org/web/packages/lss2/lss2.pdf

xtable

Coa función `xtable` podemos escribir en formato táboa de LaTex os resultados obtidos en forma de matriz mediante as simulacións. Pódese ver máis información sobre esta función en

https://cran.r-project.org/web/packages/xtable/xtable.pdf

stats

Esta librería forma parte do paquete básico de R e non fai falta cargala para poder usala. Dentro desta librería atópase a función `lm` que será de vital importancia en todo o traballo xa que a empregaremos en varias ocasións con algunha variante. No caso máis básico, para datos completos, esta función emprégase para estimar os parámetros asociados a un modelo de regresión lineal. Pódese ver máis información sobre esta función en

https://stat.ethz.ch/R-manual/R-devel/library/stats/html/lm.html

4.2. Modelo con intercepto

Estudaremos primeiro o caso do modelo de regresión con intercepto e logo veremos o caso sen intercepto xa que non todas as funcións que queremos empregar están implementados para ambos casos en R, como veremos na Sección 4.3. Ademais, para este caso

do modelo con intercepto, realizaremos dous estudos diferentes. En primeiro lugar xeraremos un modelo que cumpre as hipóteses básicas dun modelo de regresión linear, onde o erro segue unha distribución normal. En segundo lugar, realizaremos unha variante deste modelo, onde o erro vai seguir unha distribución chi-cadrado con tres grados de liberdade e observaremos as diferenzas obtidas entre estes dous escenarios para extraer conclusións sobre a sensibilidade dos métodos propostos con respecto a dita hipótese.

4.2.1. Erro con distribución normal

Comezaremos xerando en  valores do modelo

$$\textbf{Modelo 1:} \qquad Y = 1 + 2X + \varepsilon,$$

onde X representa a variable explicativa que segue unha distribución uniforme no intervalo $[0, 1]$, é dicir, $X \in U[0, 1]$, e ε representa o erro do modelo que segue unha distribución normal de media 0 e varianza σ^2, que se denota por $\varepsilon \in N\left(0, \sigma^2\right)$. Ademais, a variable de censura C seguirá unha distribución normal de varianza 1 e con diferentes medias de cara a controlar a porcentaxe de censura nos distintos escenarios considerados. Aclaramos as medias que ten que ter a variable C para os distintos casos que imos tratar nesta sección:

- Censura do 25 % (aproximadamente):
  - No caso de que $\varepsilon \in N\left(0, \frac{1}{2}\right)$, entón a variable C ten que ter unha media de 2.84 para poder acadar esta porcentaxe de censura.
  - Se $\varepsilon \in N\left(0, 1\right)$, entón a variable C ten que ter unha media de 3.04.
- Censura do 50 % (aproximadamente):
  - Se $\varepsilon \in N\left(0, \frac{1}{2}\right)$, entón a variable C ten que ter unha media de 2.
  - Se $\varepsilon \in N\left(0, 1\right)$, entón a variable C ten que ter unha media de 1.98.

Para obter estas porcentaxes de censura, realizamos unha simulación para un tamaño de mostra grande ($n = 10000$) e comprobamos as porcentaxes de censura para os diferentes escenarios considerados. Como podemos observar, se queremos aumentar a porcentaxe de censura, teremos que diminuír a media da variable de censura C.

Imos comparar catro estimadores diferentes dos parámetros $\beta_0 = 1$ e $\beta_1 = 2$ que son:

			$\widehat{\beta}_0$			$\widehat{\beta}_1$		
			Sesgo	Var	ECM	Sesgo	Var	ECM
$\sigma = 0.5$	$n = 100$	M1	0.30	4.63	4.63	−79.18	16.78	17.41
		M2	2.71	4.77	4.77	−8.98	17.57	17.58
		M3	427.81	20.17	38.47	−2995.65	110.97	1008.37
		M4	3.27	4.62	4.62	−13.59	16.60	16.62
	$n = 500$	M1	−2.64	0.97	0.97	−76.49	3.55	4.133
		M2	−2.86	0.99	0.99	2.01	3.67	3.67
		M3	400.60	4.27	20.32	−2977.70	21.64	908.31
		M4	−1.544	0.96	0.96	−1.44	3.50	3.50
	$n = 1000$	M1	1.89	0.47	0.47	−80.136	1.73	2.38
		M2	1.16	0.48	0.48	0.10	1.74	1.74
		M3	406.45	2.26	18.79	−2991.17	11.74	906.45
		M4	2.68	0.46	0.46	−4.81	1.68	1.68
$\sigma = 1$	$n = 100$	M1	−889.86	407.41	486.59	−3815.13	1261.12	2716.65
		M2	1.13	729.17	729.17	−311.70	2331.81	2341.53
		M3	26.07	650.53	650.61	−1272.151	1907.38	2069.22
		M4	−89.957	448.32	449.13	41.56	1419.34	1419.515
	$n = 500$	M1	−826.99	82.20	150.60	−4212.66	262.80	2037.46
		M2	22.74	170.84	170.90	−156.36	558.63	561.08
		M3	−173.60	178.42	181.43	−659.11	546.49	589.93
		M4	−48.70	89.96	90.20	42.94	300.13	300.32
	$n = 1000$	M1	−779.05	41.61	102.31	−4310.63	137.30	1995.44
		M2	−9.71	85.81	85.82	−54.76	273.15	273.45
		M3	−218.68	95.53	100.31	−510.42	288.39	314.43
		M4	−0.09	42.80	42.80	−33.35	142.14	142.25

Táboa 4.1: Sesgo, varianza e ECM dos estimadores obtidos (multiplicados por 10000) para o Modelo 1 a partir dos diferentes métodos M1 (estimador de mínimos cadrados ordinario), M2 (estimador proposto por Stute), M3 (estimador proposto por Stute con pesos Kaplan-Meier presuavizados) e M4 (estimador proposto por Buckley e James), sendo a porcentaxe de censura do 25 % para diferentes tamaños de mostra (denotado por n) e desviacións típicas do erro (denotado por σ).

			$\widehat{\beta}_0$			$\widehat{\beta}_1$		
			Sesgo	Var	ECM	Sesgo	Var	ECM
$\sigma = 0.5$	$n = 100$	M1	−675.90	149.58	195.26	−2135.01	620.24	1076.07
		M2	94.68	202.96	203.86	−399.08	696.21	712.14
		M3	1294.65	282.98	450.59	−5630.95	766.03	3936.79
		M4	23.63	135.25	135.30	−162.03	525.92	528.55
	$n = 500$	M1	−671.83	27.79	72.93	−2302.93	119.01	649.35
		M2	14.19	42.04	42.06	−109.78	154.20	155.41
		M3	1116.61	59.71	184.39	−5256.08	168.37	2931.01
		M4	8.42	29.27	29.28	−76.61	129.43	130.03
	$n = 1000$	M1	−627.14	14.57	53.90	−2384.88	63.80	632.57
		M2	9.75	22.36	22.37	−39.34	82.10	82.26
		M3	1077.09	33.95	149.96	−5156.54	95.99	2754.98
		M4	38.41	18.11	18.25	−106.93	84.22	85.36
$\sigma = 1$	$n = 100$	M1	−2816.86	474.36	1267.83	−5428.37	1860.84	4807.56
		M2	75.07	1361.29	1361.86	−1324.28	4525.82	4701.19
		M3	1272.32	1455.49	1617.37	−5450.49	3923.49	6894.27
		M4	−164.31	511.31	514.01	−194.59	1838.27	1842.06
	$n = 500$	M1	−2806.58	97.49	885.19	−5902.56	367.56	3851.58
		M2	19.31	462.55	462.59	−550.60	1544.08	1574.40
		M3	901.49	567.38	648.65	−4156.82	1604.62	3332.53
		M4	−101.53	104.52	105.55	−7.81	377.54	377.55
	$n = 1000$	M1	−2720.51	48.09	788.20	−6123.53	182.60	3932.37
		M2	−28.59	292.14	292.22	−260.67	936.91	943.70
		M3	729.16	361.41	414.58	−3668.39	1012.26	2357.96
		M4	−37.43	50.74	50.88	−56.47	185.97	186.29

Táboa 4.2: Sesgo, varianza e ECM dos estimadores obtidos (multiplicados por 10000) para o Modelo 1 a partir dos diferentes métodos M1 (estimador de mínimos cadrados ordinario), M2 (estimador proposto por Stute), M3 (estimador proposto por Stute con pesos Kaplan-Meier presuavizados) e M4 (estimador proposto por Buckley e James), sendo a porcentaxe de censura do 50 % para diferentes tamaños de mostra (denotado por n) e desviacións do erro (denotado por σ).

Método M1: o estimador de mínimos cadrados clásico (detallada na Sección 1.2) aplicado só sobre os datos que observamos completamente, é dicir, cando $\delta = 1$. Para aplicar este método empregaremos a función `lm` de R.

Método M2: o estimador proposto por Stute, é dicir, un estimador de mínimos cadrados ponderado con pesos Kaplan-Meier (detallado na Sección 3.1). Para aplicar este método empregaremos a función `lm` de R con argumento `weight` os pesos Kaplan-Meier calculados coa función `KMW` do paquete *condSURV*.

Método M3: unha pequena modificación do estimador proposto por Stute onde se empregan uns pesos Kaplan-Meier presuavidazados (comentado na Sección 2.5). Para aplicar este método empregaremos a función `lm` de R con argumento `weight` os pesos Kaplan-Meier presuavizados calculados coa función `PKMW` do paquete *condSURV*.

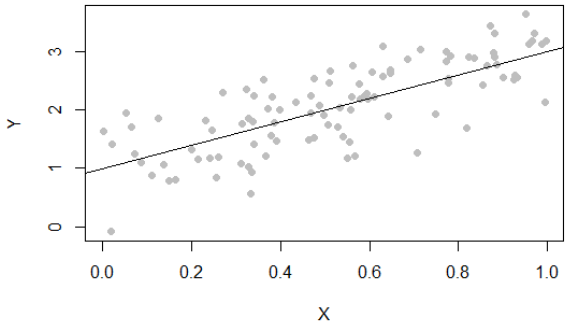
Método M4: o estimador proposto por Buckley e James (detallado na Sección 3.3). Para aplicar este método empregaremos a función `bf` do paquete *rms* de R.

Para comparar os métodos anteriores calcularemos o sesgo, a varianza e o ECM de cada estimador. Os resultados pódense ver nas Táboas 4.1 e 4.2 que teñen asociadas unha porcentaxe do 25 % e 50 % de censura, respectivamente. Nestas táboas atopamos os resultados multiplicados por 10000 para poder comparar ben os ECM. Así evitamos a aparición de valores 0.000 ao aproximar os resultados. En cada táboa calculamos as medidas resumo para diferentes desviacións típicas do erro e diferentes tamaños de mostra onde en cada escenario se realizaron 1000 réplicas Monte Carlo.

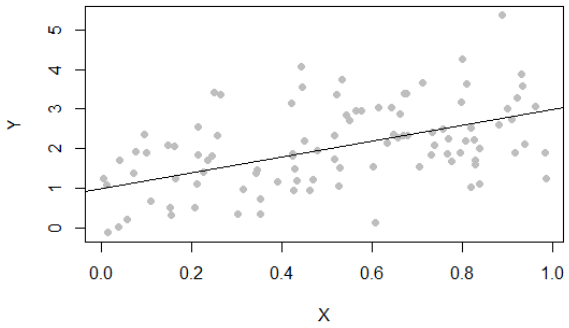
O código empregado para levar a cabo este estudo de simulación atópase no Anexo A deste traballo. En canto á programación, destacamos que hai que redefinir o maior dato da variable resposta observada como non censurado para que o estimador de Stute, M2, sexa consistente (problema derivado da consistencia do estimador de Kaplan-Meier). Ademais, o método de Buckley e James, M4, non sempre converxe, polo que temos que eliminar as iteracións que non converxan para obter uns resultados que poidamos comparar co resto de métodos.

En primeiro lugar, se observamos as Táboas 4.1 e 4.5 en conxunto, podemos sacar as seguintes conclusións xerais:

- O ECM dos estimadores aumenta canto máis grande sexa a varianza do erro, o cal é lóxico ao haber máis dispersión nos datos como se pode ver na Figura 4.1.



(a) Desviación típica de 0.5.



(b) Desviación típica de 1.

Figura 4.1: Representación gráfica dunha mostra de tamaño $n = 100$ do Modelo 1 xunto coa recta de regresión teórica para diferentes desviacións típicas da distribución do erro.

- O ECM diminúe conforme aumentamos o tamaño de mostra, o cal tamén é de esperar xa que ao ter un tamaño de mostra maior, temos máis información.
- O ECM aumenta cando aumenta a porcentaxe de censura. O ECM é menor en todos os métodos para o caso de censura do 25 % se os comparamos co obtido nos casos de censura de 50 %. Este feito débese a que ao aumentar a censura imos ter menos información e as estimacións serán menos precisas.
- Para todos os métodos considerados resulta mellor a estimación do intercepto que a da pendente, pois sempre ten un ECM máis baixo.

A modo de exemplo, imos observar os datos obtidos para o intercepto e a pendente cando a desviación típica do erro é 0.5 e ímonos fixar nas Táboas 4.1 e 4.2. En ambos ca-

sos, o método M3 é o que peor resultados nos proporciona, pois os seus ECM son os máis elevados. Este feito resulta curioso posto que os pesos Kaplan-Meier presuavizados proporcionan mellores resultados que os clásicos pesos Kaplan-Meier na estimación da función de distribución baixo censura. Porén isto non se observa no contexto da regresión censurada onde os mellores resultados son os asociados aos clásicos pesos Kaplan-Meier. Ademais, pese a que a primeira vista poderíamos pensar que o método M1 non ía proporcionar bos estimadores dado que estamos tendo en conta só os datos que observamos completamente, tamén temos que ter en conta que estamos considerando unha censura dun 25 % ou dun 50 % e unha desviación do erro de 0.5. Así, para un tamaño de mostra de $n = 1000$, contamos con en torno a 750 ou 500 observacións respectivamente e é lóxico que este método sexa capaz de aproximar ben o modelo tendo en conta a distribución do erro.

Imos ver que ocorre cando a desviación típica do erro é 1. No caso de contar cunha porcentaxe de censura do 25 %, na Táboa 4.1, observamos que a pendente estímase peor sempre mediante M1. Destacamos que para un tamaño de 100, M3 estima mellor tanto β_0 como β_1 que M2. En canto ao intercepto, para este escenario, non obtemos un resultado unánime con respecto a cal é o peor método para estimalo. Se temos en conta agora unha censura do 50 %, na Táboa 4.2 podemos ver que neste caso o peor método é M1. Simplemente facemos unha excepción para o tamaño de mostra de 100, onde o peor volvería ser M3. Estes resultados poñen de manifesto a utilidade de presentar métodos de estimación específicos para o contexto de datos censurados.

Como conclusión xeral desta simulación, observamos que, en termos de ECM, o que da mellores resultados sempre é o asociado ao método M4 xa que sempre se observan menores ECM para todos os casos. Este resultado coincide co que tiñamos pensado atoparnos antes de iniciar a simulación, pois o estimador proposto por Buckley e James é o máis fiable destes catro métodos. O seguinte método que estima mellor os parámetros, no caso da pendente, sería M2 e finalmente teríamos M1 e M3. Para o caso do intercepto, o segundo mellor método non está moi ben definido xa que para cada escenario contamos con diferentes conclusións.

4.2.2. Erro con distribución chi-cadrado

Na Sección 4.2.1 observamos o caso no que o erro segue unha distribución normal, que é unha das hipóteses do modelo de regresión linear simple. Que ocorrería se consideramos outra distribución para o erro que non sexa normal? Nesta sección imos estudar este caso para salientar a importancia de comprobar as hipóteses para realizar unha boa Inferencia

			$\widehat{\beta}_0$			$\widehat{\beta}_1$		
			Sesgo	Var	ECM	Sesgo	Var	ECM
Censura: 25 %	$n = 100$	M1	2.225	0.091	5.043	−0.494	0.271	0.516
		M2	2.777	0.795	8.508	−0.202	2.728	2.769
		M3	2.784	0.698	8.448	−0.219	2.362	2.409
		M4	2.760	0.159	7.775	−0.002	0.480	0.480
	$n = 500$	M1	2.212	0.017	4.912	−0.531	0.048	0.330
		M2	2.815	0.548	8.471	−0.116	1.913	1.926
		M3	2.813	0.754	8.665	−0.109	2.653	2.665
		M4	2.825	0.034	8.016	−0.006	0.098	0.099
	$n = 1000$	M1	2.207	0.009	4.879	−0.531	0.024	0.306
		M2	2.845	0.441	8.533	−0.120	1.558	1.573
		M3	2.844	0.731	8.821	−0.101	2.601	2.612
		M4	2.845	0.018	8.112	−0.005	0.048	0.048
Censura: 50 %	$n = 100$	M1	1.718	0.074	3.025	−0.548	0.242	0.542
		M2	2.511	1.104	7.410	−0.323	3.816	3.920
		M3	2.678	0.895	8.069	−0.673	2.961	3.413
		M4	2.543	0.131	6.598	−0.006	0.366	0.367
	$n = 500$	M1	1.718	0.014	2.964	−0.612	0.040	0.414
		M2	2.613	0.795	7.620	−0.221	2.721	2.770
		M3	2.740	0.933	8.443	−0.511	3.115	3.376
		M4	2.654	0.034	7.077	−0.013	0.081	0.081
	$n = 1000$	M1	1.713	0.006	2.942	−0.617	0.018	0.399
		M2	2.684	0.719	7.924	−0.262	2.496	2.564
		M3	2.804	0.926	8.786	−0.528	3.166	3.445
		M4	2.685	0.018	7.225	−0.006	0.037	0.037

Táboa 4.3: Sesgo, varianza e ECM dos estimadores obtidos para o Modelo 1B a partir dos diferentes métodos M1 (estimador de mínimos cadrados ordinario), M2 (estimador proposto por Stute), M3 (estimador proposto por Stute con pesos Kaplan-Meier presuavizados) e M4 (estimador proposto por Buckley e James), sendo unha porcentaxe de censura do 25 % e 50 % para diferentes tamaños de mostra (denotado por n).

Estatística.

Comezaremos xerando en  valores do modelo

$$\textbf{Modelo 1B:} \qquad Y = 1 + 2X + \varepsilon,$$

onde X representa a variable explicativa que segue unha distribución uniforme no intervalo $[0, 1]$, é dicir, $X \in U[0, 1]$, e ε representa o erro do modelo que segue unha distribución chi-cadrado[1] con tres grados de liberdade, é dicir, $\varepsilon \in \chi^2_3$.

Ademais, a variable de censura C seguirá unha distribución normal de varianza 1 e con diferentes medias de cara a controlar a porcentaxe de censura nos distintos escenarios considerados. Aclaramos as medias que ten que ter a variable C para os distintos casos que imos tratar nesta sección:

- Censura do 25 % (aproximadamente):
  - A variable C ten que ter unha media de 6.34 para poder acadar esta porcentaxe de censura.
- Censura do 50 % (aproximadamente):
  - A variable C ten que ter unha media de 4.52 para poder acadar esta porcentaxe de censura.

Calcularemos o sesgo, a varianza e o ECM de cada estimador para os diferentes métodos M1 (estimador de mínimos cadrados ordinario), M2 (estimador proposto por Stute), M3 (estimador proposto por Stute con pesos Kaplan-Meier presuavizados) e M4 (estimador proposto por Buckley e James) explicados na Sección 4.2.1. Os resultados pódense ver na Táboa 4.3 onde consideramos diferentes tamaños de mostra (denotado por n) e en cada escenario se realizaron 1000 réplicas Monte Carlo. A única diferenza salientable en canto a programación é a distribución do erro, que neste caso é unha chi-cadrado de tres grados de liberdade. Polo tanto, para xerar a variable do erro, empregaremos o comando `rchisq`.

Na Táboa 4.3 podemos observar os datos obtidos e, se nos fixamos nos ECM, podemos ver que son valores máis altos do habitual. Isto quere dicir que as estimacións non son boas.

[1] Sexan $Z_1, \ldots, Z_m$ variables aleatorias normais estándar independentes. Diremos que a variable aleatoria

$$X = Z_1^2 + \cdots + Z_m^2$$

segue unha distribución chi-cadrado con m grados de liberdade, onde χ^2_m é a notación para a distribución chi-cadrado e o subíndice m representa os grados de liberdade.

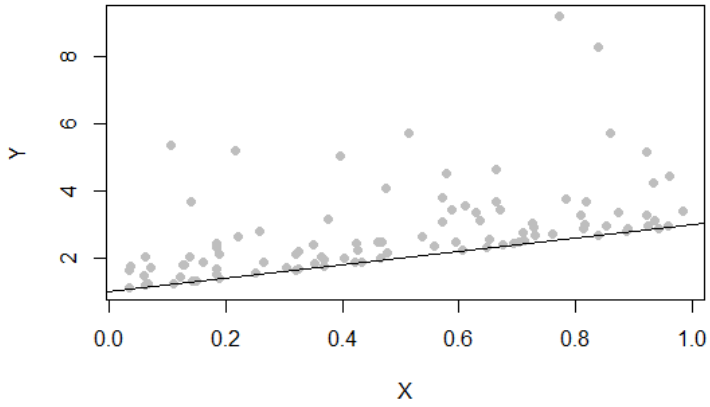


Figura 4.2: Representación gráfica dunha mostra de tamaño $n = 100$ do Modelo 1B xunto coa recta de regresión teórica.

Sen embargo, o método M4 é capaz de obter outra vez as mellores estimacións, aínda que, ao non cumprir a hipótese de normalidade dos erros, estes resultados non os poderemos empregar para facer inferencia, pois non obteremos resultados fiables.

Se observamos a Figura 4.2, podemos ver como a nube de puntos se distribúe en torno á recta de regresión poñendo de manifesto a asimetría da distribución do erro. Con este exemplo ilustramos a importancia de validar as hipóteses dun modelo de regresión lineal.

4.3. Modelo sen intercepto

4.3.1. Erro con distribución normal

Consideremos agora o modelo

$$\textbf{Modelo 2:} \qquad Y = 2X + \varepsilon,$$

onde X representa a variable explicativa que segue unha distribución uniforme no intervalo [0,1], é dicir, $X \in U[0,1]$ e ε representa o erro do modelo que segue unha distribución normal de media 0 e varianza σ^2, que se denota por $\varepsilon \in N\left(0, \sigma^2\right)$. Ademais, como no caso anterior, a variable de censura C seguirá unha distribución normal de varianza 1 e con diferentes medias de cara a manter constante a porcentaxe de censura nos distintos

		$\sigma = 0.5$			$\sigma = 1$		
	Métodos	Sesgo	Varianza	ECM	Sesgo	Varianza	ECM
$n = 50$	M1	−0.167	0.022	0.050	−0.523	0.069	0.343
	M2	−0.011	0.023	0.023	−0.044	0.080	0.082
	M3	−0.200	0.019	0.059	−0.149	0.062	0.084
	M4	−0.017	0.080	0.080	−0.031	0.286	0.287
	M5	−0.004	0.021	0.021	−0.012	0.076	0.076
	M6	−0.015	0.081	0.081	−0.025	0.292	0.293
$n = 100$	M1	−0.171	0.010	0.040	−0.539	0.032	0.323
	M2	−0.002	0.011	0.011	−0.019	0.041	0.041
	M3	−0.191	0.009	0.046	−0.116	0.032	0.046
	M4	0.010	0.039	0.039	0.021	0.144	0.144
	M5	0.002	0.010	0.010	0.003	0.036	0.036
	M6	0.011	0.039	0.039	0.022	0.144	0.144
$n = 200$	M1	−0.176	0.006	0.037	−0.553	0.017	0.324
	M2	0.002	0.005	0.005	−0.012	0.021	0.021
	M3	−0.192	0.005	0.042	−0.104	0.017	0.028
	M4	0.003	0.019	0.019	0.002	0.072	0.072
	M5	0.004	0.005	0.005	0.004	0.019	0.019
	M6	0.005	0.019	0.019	0.003	0.072	0.072

Táboa 4.4: Sesgo, varianza e ECM dos estimadores da pendente obtidos para o Modelo 2 a partir dos diferentes métodos M1 (estimador de mínimos cadrados ordinario), M2 (estimador proposto por Stute), M3 (estimador proposto por Stute con pesos Kaplan-Meier presuavizados) e M4 (estimador proposto por Buckley e James), M5 (estimador proposto por Miller) e M6 (estimador proposto por Jin, Lin e Ying) sendo unha porcentaxe de censura do 25 % para diferentes tamaños de mostra (denotado por n) e desviacións do erro (denotado por σ).

		$\sigma = 0.5$			$\sigma = 1$		
	Métodos	Sesgo	Varianza	ECM	Sesgo	Varianza	ECM
$n = 50$	M1	−0.315	0.037	0.136	−0.967	0.104	1.039
	M2	−0.034	0.041	0.042	−0.135	0.127	0.145
	M3	−0.390	0.026	0.178	−0.386	0.078	0.227
	M4	−0.024	0.105	0.106	−0.035	0.376	0.377
	M5	−0.010	0.034	0.034	−0.025	0.115	0.116
	M6	−0.012	0.109	0.109	−0.019	0.380	0.380
$n = 100$	M1	−0.334	0.019	0.131	−1.012	0.052	1.076
	M2	−0.014	0.023	0.023	−0.084	0.068	0.075
	M3	−0.373	0.014	0.153	−0.339	0.043	0.157
	M4	0.002	0.057	0.057	0.018	0.193	0.193
	M5	−0.002	0.018	0.018	−0.006	0.059	0.059
	M6	0.010	0.057	0.057	0.026	0.193	0.194
$n = 200$	M1	−0.335	0.010	0.123	−1.031	0.027	1.090
	M2	−0.005	0.011	0.011	−0.056	0.035	0.038
	M3	−0.369	0.007	0.143	−0.299	0.025	0.114
	M4	−0.006	0.028	0.028	−0.005	0.093	0.093
	M5	0.005	0.009	0.009	0.005	0.029	0.029
	M6	0.001	0.028	0.028	0.003	0.092	0.092

Táboa 4.5: Sesgo, varianza e ECM dos estimadores da pendente obtidos para o Modelo 2 a partir dos diferentes métodos M1 (estimador de mínimos cadrados ordinario), M2 (estimador proposto por Stute), M3 (estimador proposto por Stute con pesos Kaplan-Meier presuavizados) e M4 (estimador proposto por Buckley e James), M5 (estimador proposto por Miller) e M6 (estimador proposto por Jin, Lin e Ying) sendo unha porcentaxe de censura do 50 % para diferentes tamaños de mostra (denotado por n) e desviacións do erro (denotado por σ).

escenarios considerados. Aclaramos as medias que ten que ter a variable C para os distintos casos que imos tratar:

- Censura do 25 % (aproximadamente):
  - Se $\varepsilon \in N\left(0, \frac{1}{2}\right)$, entón a variable C ten que ter unha media de 1.85.
  - Se $\varepsilon \in N\left(0, 1\right)$, entón a variable C ten que ter unha media de 2.
- Censura do 50 %(aproximadamente):
  - Se $\varepsilon \in N\left(0, \frac{1}{2}\right)$, entón a variable C ten que ter unha media de 1.
  - Se $\varepsilon \in N\left(0, 1\right)$, entón a variable C ten que ter unha media de 1.

Para obter estes resultados, realizamos unha simulación para un tamaño de mostra grande ($n = 10000$) e comprobamos as censuras para os datos indicados. Como podemos observar, se queremos aumentar a porcentaxe de censura, teremos que diminuír a media da variable de censura C.

Imos comparar seis estimadores diferentes con parámetro $\beta_1 = 2$. Os métodos empregados serán os catro métodos que empregamos na Sección 4.2.1 no caso do modelo con intercepto, M1 (estimador de mínimos cadrados ordinario), M2 (estimador proposto por Stute), M3 (estimador proposto por Stute con pesos Kaplan-Meier presuavizados) e M4 (estimador proposto por Buckley e James), e os dous seguintes:

Método M5: o estimador proposto por Miller (detallado na Sección 3.2). Para aplicar este método empregaremos a función `CensReg.SMN`, pero o argumento `status` sería $1 - \delta$ (o contrario da nosa notación). Para nós, recordemos, que $\delta = 0$ é un dato censurado e $\delta = 1$ é un dato non censurado.

Método M6: o estimador proposto por Jin, Lin e Ying (detallado na Sección 3.4). Para aplicar este método empregaremos a función `lss`.

A razón pola que separamos o modelo con intercepto do modelo sen intercepto é pola forma na que están programadas as funcións `CensReg.SMN` e `lss` en R. Estas dúas funcións, que se corresponden cos métodos M5 e M6, so estiman a pendente e consideran so o caso dun modelo sen intercepto.

Para levar a cabo isto, como na Sección 4.2.1, empregaremos o programa estatístico R e calcularemos o sesgo, a varianza e o ECM de cada estimador. Os resultados pódense

ver nas Táboas 4.4 e 4.5 que foron feitas cun 25 % e 50 % de porcentaxe de censura respectivamente. En cada táboa calculamos as medidas resumo para diferentes desviacións típicas do erro (denotado por σ) e diferentes tamaños de mostra (denotado por n) onde en cada escenario se realizaron 1000 réplicas Monte Carlo. Como diferenza do feito nas simulacións do Modelo 1, debido ao elevado tempo de computación dos métodos M5 e M6, realizaremos esta simulación para tamaños de mostra máis pequenos, sendo $n = 50$, $n = 100$ e $n = 200$.

O código empregado para levar a cabo esta simulación atópase no Anexo A deste traballo. En canto á programación, destacamos que o Método M4, é dicir, o método proposto por Buckley e James, tamén está estimando o intercepto[2] e polo tanto está en "desvantaxe" co resto dos métodos. Non obstante, non observamos unha mala estimación en xeral, polo que se decidiu deixalo nesta simulación. Os métodos M1, M2 e M3 modificámolos minimamente para que só estimasen a pendente e seguimos empregando a función `lm`.

Se observamos as Táboas 4.4 e 4.5, a primeira vista podemos observar as mesmas conclusión xerais básicas mencionadas na Sección 4.2.1, é dicir, observamos que o ECM diminúe cando aumentamos o tamaño de mostra, que aumenta se aumentamos a porcentaxe de censura e que tamén se ve incrementado para valores de desviación típica do erro máis altos. Ademais, observamos tamén comportamentos similares entre os resultados obtidos con desviación típica de 0.5 para unha porcentaxe de censura do 25 % e para unha porcentaxe de censura do 50 %. Así mesmo, tamén obtemos as mesmas conclusións para o caso cando a desviación típica toma valor 1. Vexamos estas semellanzas con máis detalle.

Imos observar os datos obtidos cando $\sigma = 0.5$ para as dúas posibles porcentaxes de censura: 25 % e 50 %. Neste escenario, o método que peor estima a pendente é M3, exceptuando no caso de 25 % de censura para tamaño de mostra $n = 50$, onde observamos, sorprendentemente, que o peor método é M6. Ademais, para tamaños $n = 100$ e $n = 200$, para ambas porcentaxes de censura, observamos que os ECM de M4 e de M6 son iguais. Este resultado parece confirmar o que estudamos na Sección 3.4, xa que cando aumentamos o tamaño de mostra, o estimador proposto por Jin, Lin e Ying (M6) aproxímase ao estimador proposto por Buckley e James (M4). Ademais, exceptuando o caso que xa comentamos de censura do 25 % e tamaño $n = 50$, para o resto de escenarios os método M4 e M6 (posto que teñen os mesmos ECM), estarían en terceira posición para a ser o mellor método, que so é superado por M5 e M2.

Imos ver agora que ocorre cando a desviación típica do erro é 1. Neste caso, o peor método que estima a pendente é M1 para ambas porcentaxes de censura de 25 % e 50 %. No caso de censura do 25 % observamos tamén a aproximación das estimacións dos métodos

[2] A función `bj` non permite a opción de eliminar o intercepto.

M4 e M6 para os tamaños $n = 100$ e $n = 200$. Destacamos que para este valor de desviación típica, o método M3 é o terceiro mellor método, exceptuando o caso de 50 % de censura para $n = 200$, onde este método M3 se atoparía en quinta posición. Neste último caso particular observamos que en terceira posición se atopa M6, que ten un ECM menor que M4 e esta é a única ocasión no que acontece isto xa que nos casos restantes ou ben M4 ten un ECM máis baixo que M6 ou ben estes erros son iguais.

Como conclusión xeral desta simulación, observamos que, se nos fixamos nos valores dos ECM obtidos, o que da mellores resultados sempre é o asociado ao método M5, é dicir, o estimador proposto por Miller. En segundo lugar, o seguinte mellor método observamos tamén sempre que é o método M2, é dicir, o estimador proposto por Stute onde se empregan os pesos de Kaplan-Meier. Despois, o resto de métodos depende da desviación típica do erro que esteamos considerando e, nalgún caso puntual, tamén dependería da censura considerada, como se comentou con anterioridade. Podemos sacar como conclusión que para estimar modelos cunha variabilidade máis alta, sen ter en conta M5 e M2, o método que vai dar mellores resultados será o método M3. Se pola contra queremos estimar os parámetros dun modelo con pouca variabilidade, entón podemos facer uso do modelo M4 ou M6, indistintamente.

4.3.2. Erro con distribución chi-cadrado

Consideremos agora o modelo

$$\textbf{Modelo 2B:} \qquad Y = 2X + \varepsilon,$$

onde X representa a variable explicativa que segue unha distribución uniforme no intervalo [0,1], é dicir, $X \in U[0,1]$ e ε representa o erro do modelo que segue unha distribución chi-cadrado con tres grados de liberdade, que se denota por $\varepsilon \in \chi^2_3$. Ademais, como nos casos anteriores, a variable de censura C seguirá unha distribución normal de varianza 1 e con diferentes medias de cara a manter constante a porcentaxe de censura nos distintos escenarios considerados. Aclaramos as medias que ten que ter a variable C para os distintos casos que imos tratar nesta sección:

- Censura do 25 % (aproximadamente):
  - A variable C ten que ter unha media de 5.31 para poder acadar esta porcentaxe de censura.
- Censura do 50 % (aproximadamente):

		Censura 25 %			Censura 50 %		
	Métodos	Sesgo	Varianza	ECM	Sesgo	Varianza	ECM
$n = 50$	M1	3.027	0.173	9.336	2.351	0.186	5.715
	M2	4.073	0.368	16.954	3.716	0.676	14.485
	M3	3.992	0.290	16.223	3.385	0.338	11.799
	M4	−0.057	0.931	0.934	−0.065	0.725	0.730
	M5	4.331	0.381	19.140	4.546	0.666	21.336
	M6	−0.051	0.934	0.937	−0.056	0.727	0.730
$n = 100$	M1	2.931	0.082	8.675	2.238	0.081	5.091
	M2	4.091	0.225	16.965	3.744	0.403	14.420
	M3	4.059	0.201	16.674	3.487	0.246	12.408
	M4	−0.047	0.463	0.466	−0.027	0.364	0.365
	M5	4.315	0.201	18.820	4.514	0.330	20.704
	M6	−0.045	0.464	0.466	−0.022	0.369	0.369
$n = 200$	M1	2.896	0.040	8.428	2.187	0.040	4.825
	M2	4.128	0.133	17.173	3.782	0.260	14.562
	M3	4.114	0.128	17.051	3.565	0.179	12.889
	M4	−0.041	0.246	0.248	−0.028	0.196	0.197
	M5	4.309	0.096	18.665	4.532	0.168	20.708
	M6	−0.040	0.247	0.248	−0.024	0.196	0.196

Táboa 4.6: Sesgo, varianza e ECM dos estimadores da pendente obtidos para o Modelo 2B a partir dos diferentes métodos M1 (estimador de mínimos cadrados ordinario), M2 (estimador proposto por Stute), M3 (estimador proposto por Stute con pesos Kaplan-Meier presuavizados) e M4 (estimador proposto por Buckley e James), M5 (estimador proposto por Miller) e M6 (estimador proposto por Jin, Lin e Ying) sendo unha porcentaxe de censura do 25 % e 50 % para diferentes tamaños de mostra (denotado por n).

- A variable C ten que ter unha media de 3.5 para poder acadar esta porcentaxe de censura.

Para obter estes resultados, realizamos unha simulación para un tamaño de mostra grande ($n = 10000$) e comprobamos as censuras para os datos indicados. Como podemos observar, se queremos aumentar a porcentaxe de censura, teremos que diminuír a media da variable de censura C.

Calcularemos o sesgo, a varianza e o ECM de cada estimador para os diferentes métodos M1 (mínimos cadrados usual), M2 (método de Stute), M3 (método de Stute con pesos Kaplan-Meier presuavizados), M4 (método de Buckley e James), M5 (método de Miller) e M6 (método de Jin, Lin e Ying) explicados na Sección 4.3.1. Os resultados pódense ver na táboa 4.6 onde consideramos diferentes tamaños de mostra e en cada escenario se realizaron 1000 réplicas Monte Carlo.

Na Táboa 4.6 podemos observar os datos obtidos e, se nos fixamos nos ECM, podemos ver que son valores máis altos que os obtidos para o erro normal. Observamos unha gran diferenza coas conclusións que obtivemos na Sección 4.3.1 onde o erro seguía unha distribución normal xa que o método M5 era o mellor e neste caso non ocorre iso. O método M4-M6 volve ser o que mellor estimacións proporciona, seguido de M1, M3, M2 e finalmente M5.

Concluímos desta forma que pese a que M5 é mellor no caso de ter un modelo sen intercepto cando o seu erro ten distribución normal, M4-M6 son métodos máis robustos que funcionan ben aínda que se incumpra algunha das hipóteses básicas do modelo de regresión lineal simple.

Capítulo 5

Aplicación a datos reais

Neste capítulo imos aplicar os métodos estudados ao longo dos Capítulos 2 e 3 a unha base de datos reais. Dito estudo completará as observacións feitas no Capítulo 4. Empregaremos para isto a base de datos UIS procedente do paquete `quantreg` do programa estatístico R. Pódese ver máis información sobre este paquete en

https://cran.r-project.org/web/packages/quantreg/quantreg.pdf

5.1. Base de datos UIS

A base de datos UIS[1] foi elaborada polas médicas Jane McCusker, Carol Bigelow e Anne Stoddatd no marco dun estudo realizado na Unidade de Investigación da Universidade de Massachusetts. O estudo consistía nun proxecto composto por dous ensaios clínicos aleatorios sobre o tratamento abusivo de drogas en residencias que tivo unha duración de cinco anos, dende o ano 1989 ata o ano 1994. O propósito deste estudo era o de comparar os programas de tratamento de diferentes duracións deseñados para reducir o abuso de drogas e para prever o alto risco de VIH (virus da inmunodeficiencia humana). A continuación detállase o alcance de ambos ensaios clínicos aos cales nos referiremos como A e B.

[1] University of Massachusetts Aids Research Unit (UMARU) Impact Study (UIS).

No ensaio A formaron parte 444 persoas e consistiu nunha comparación de terapias que se modificaron en comunidades onde se incorporou un programa de educación sobre a saúde e sobre a prevención de recaídas. Este estudo tivo unha duración de tres a seis meses. Ademais, ás/aos participantes ensinóuselle a recoñecer as situacións que supoñen un alto risco que poden ocasionar unha posible recaída e tamén se lles ensinaron habilidades que lle permitían afrontar estas situacións complicadas sen ter que facer uso das drogas.

No ensaio B participaron 184 persoas que recibían un programa terapéutico ou ben de seis meses ou de doce meses de duración. Este programa involucraba un estilo de vida perfectamente estruturado nun entorno de vida comunal onde os membros compartían diferentes aspectos da súa vida mediante un vínculo.

A base de datos UIS facilítanos información sobre ambos ensaios. Máis concretamente, na Táboa 5.1 atópase o conxunto de variables recollidas ao longo do estudo.

A variable `TIME` considerarase como a variable resposta na análise estatística que imos realizar e defínese como o número de días dende que se realiza a admisión do/a individuo/a en calquera dos dous centros posibles ata que esa persoa teña unha recaída no uso das drogas. Nótese que se trata dunha variable censurada onde a variable `CENSOR` será a variable indicadora da censura que se produce como consecuencia dunha perda do seguimento dun/-dunha doente. Neste estudo, que conta cun 19.3 % de censura, tamén se considera que se unha persoa sae do ensaio e, polo tanto, se perde o seguimento, é debido a unha recaída no uso das drogas. Poderíamos realizar diferentes estudos tendo en conta as posibles variables explicativas que poden ser significativas para explicar o comportamento da variable resposta. Ademais, contamos cunha serie de variables categóricas que nos servirán para facer subgrupos e observar o comportamento dependendo, por exemplo, da raza do/a doente ou do tipo de droga consumida antes de entrar ao ensaio. As posibles **variables explicativas** consideradas serán:

`AGE, BECK, NDT, LEN.T`

e as variables categóricas que teremos en conta para os nosos subgrupos serán:

`HERCOC, IV, RACE, TREAT e SITE.`

Variable	Descrición	Códigos/Valores
ID	Código de identificación	1-628
AGE	Idade de inscrición	Anos
BECKTOTA	Puntuación de depresión na admisión	$0.000 - 54.000$
HERCOC	Uso de heroína ou cocaína durante meses antes da admisión	1=Heroína e cocaína 2=So heroína 3= So cocaína 4=Nin cocaína nin heroína
IVHX	Historia do uso da droga	1=Nunca 2=Previamente 3=Recentemente
NDRUGTX	Número de tratamentos anteriores de drogas	$0 - 40$
RACE	Raza do/a individuo/a	0=Branca 1=Non branca
TREAT	Tratamento asignado	0=Curto 1=Largo
SITE	Sitio do tratamento	0=A 1=B
LEN.T	Período de estancia no tratamento	Días
TIME	Tempo ata a recaída	Días
CENSOR	Evento de perda do seguimento ou recaída no uso das drogas	1=Volver ás drogas ou perda do seguimento 0=Outro caso
Y	Logaritmo da variable TIME	
FRAC	Lonxitude do tratamento fraccionado	LEN.T/90, tratamento curto LEN.T/180, tratamento longo
IV3	Uso recente de drogas	1=Si 0=Non

Táboa 5.1: Explicación das variables procedentes da base de datos UIS.

5.2. Análise descritiva previa

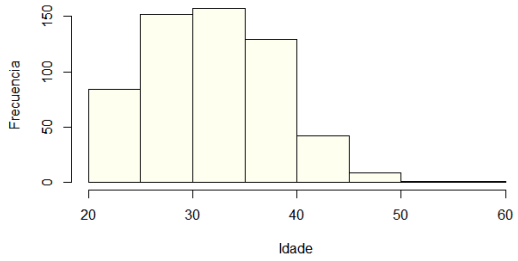
Pode ser interesante realizar un estudo inicial básico para observar as características dos datos que imos analizar máis adiante. Na Figura 5.1 podemos observar os histogramas das posibles variables explicativas que poderíamos incluír nun modelo de regresión para explicar a variable `TIME`. Ademais destes gráficos, tamén podemos calcular medidas características como a media, o mínimo ou o máximo de cada unha destas variables que obtemos mediante a función `summary` de  . Así, observamos que a idade media dos/as individuos/as que participaron no estudo é de 32.38 anos mentres que as idades están comprendidas entre os 20 e os 56 anos. No caso da variable `BECK`, obtemos que o índice de depresión entre os e as participantes ten unha media de 17.37 e os seus valores atópanse entre 0 e 54. En canto ao número de tratamentos anteriores de droga, `NDT`, observamos que o número medio é de 4.543 tratamentos e atopamos individuos/as con 0 tratamentos e tamén con 40 tratamentos, sendo este o máximo desta variable. Por último, destacamos que o número medio de días de estancia no tratamento, variable `LEN.T`, é de 100.8 días, sendo o mínimo 3 e o máximo de estadía 400 días.

En canto ás variables categóricas, podemos definir subgrupos con certas características a partir da base de datos orixinal. En primeiro lugar, se comezamos estudando a variable `HC`, podemos ver como 104 persoas consumían heroína e cocaína, 107 so heroína, 172 so cocaína e 192 ningunha destas dúas drogas. Por outra banda, no estudo participaron 430 individuos/as de raza branca e 145 de raza non branca. Ademais, no tratamento curto formaron parte 289 persoas e no tratamento longo 286. Finalmente, tal como se explicou ao comezo, 400 persoas participaron no ensaio no lugar A e 175 no lugar B.

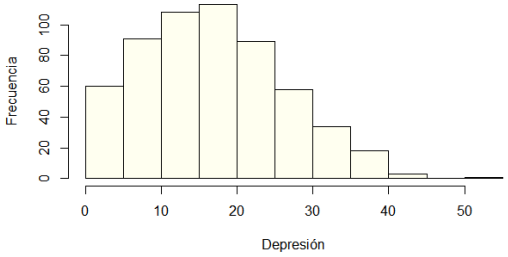
5.3. Estimación dun modelo de regresión

Realizaremos un estudo para estimar o intercepto e a pendente de diferentes modelos propostos a partir da base de datos que acabamos de introducir. Parece interesante estudar como afecta a idade ao tempo que pasa ata que se recae no consumo de drogas e comprobar como é a relación entre ambas variables. Para isto empregaremos un modelo de regresión lineal simple entre a variable explicativa `AGE` e a variable resposta `TIME` descritas na Táboa 5.1 que podemos escribir como:

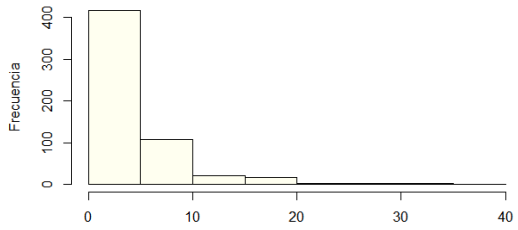
$$\texttt{TIME} = \beta_0 + \beta_1 \texttt{AGE} + \varepsilon, \tag{5.1}$$



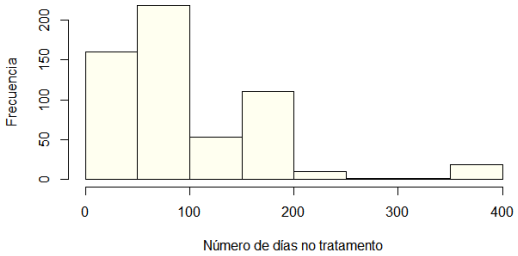
(a) Histograma da variable AGE.



(b) Histograma da variable BECK.



(c) Histograma da variable NDT.



(d) Histograma da variable LEN.T.

Figura 5.1: Histogramas das posibles variables explicativas consideradas para explicar o comportamento da variable TIME.

sendo ε o erro que segue unha distribución normal e neste caso a variable resposta é o tempo que pasa ata a recaída nas drogas, é dicir, TIME, e a variable explicativa será a idade, é dicir, AGE. Ademais, cos coñecementos obtidos durante a realización do estudo de simulación, para estimar o intercepto e a pendente sabemos que o método que nos proporciona mellores resultados é M4, é dicir, empregaremos o estimador proposto por Buckley e James. Por outra banda, se obtemos resultados que nos indiquen que o intercepto non é significativo, entón sabemos tamén do estudo de simulación que o mellor método que podemos empregar é M5, é dicir, o estimador proposto por Miller. Estes métodos e a súa programación está explicada con máis detalle nas Seccións 4.2 e 4.3, respectivamente.

Mostramos a continuación o código empregado en  así como as saídas resultantes para describir a relación entre a idade e o tempo ata a recaída:

```
>bj(Surv(TIME,CENSOR)~AGE,link="identity",x=TRUE, y=TRUE)

Buckley-James Censored Data Regression
```

```
bj(formula = Surv(TIME, CENSOR) ~ AGE, link = "identity", x = TRUE,
y = TRUE)

Discrimination
Indexes
Obs    575    Regression d.f.1    g       33.396
Events 464    sigma124.2258
d.f.      462

     Coef    S.E.  Wald Z Pr(>|Z|)
Intercept 183.5572 31.2655 5.87   <0.0001
AGE         4.7533  0.9541 4.98   <0.0001
```

Observación 5.1. Para comprobar que os valores estimados son significativos, recordemos que temos que fixarnos na columna de $Pr(> |Z|)$, onde se obtén o nivel crítico para o contraste de que o coeficiente é cero. Se o nivel crítico é menor que os niveis de significación habituais (10 %, 5 % ou 1 %) entón diremos que dito coeficiente é significativamente distinto de cero. Por outra banda, no caso de que non sexa significativo, entón poderemos considerar que dito parámetro é cero, e polo tanto, non aporta nada ao modelo de regresión. Se obtemos que a pendente non é significativa, estamos dicindo que as variables non estarían relacionadas.

Nesta saída observamos que tanto o intercepto como a pendente son significativamente distintas de cero e teñen valores de 188.557 días e 4.753 días/anos respectivamente. Na Figura 5.2 mostramos o diagrama de dispersión xunto co axuste obtido e observamos unha relación lineal crecente entre ambas variables.

Ademais, tamén podemos pensar en axustar o modelo (5.1) para os diferentes subgrupos que determinan as variables categóricas proporcionadas pola base de datos e, como xa introducimos antes, empregaremos a raza, variable `RACE`, e o tipo de droga consumida, variable `HC`.

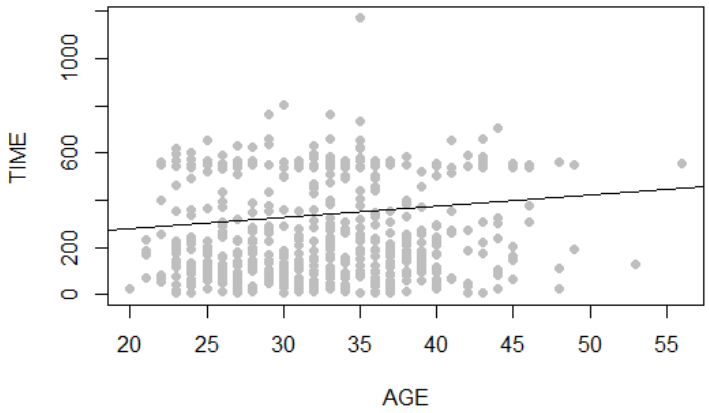


Figura 5.2: Representación da idade dos/as participantes no ensaio clínico fronte ao tempo ata a recaída xunto co modelo (5.1) axustado.

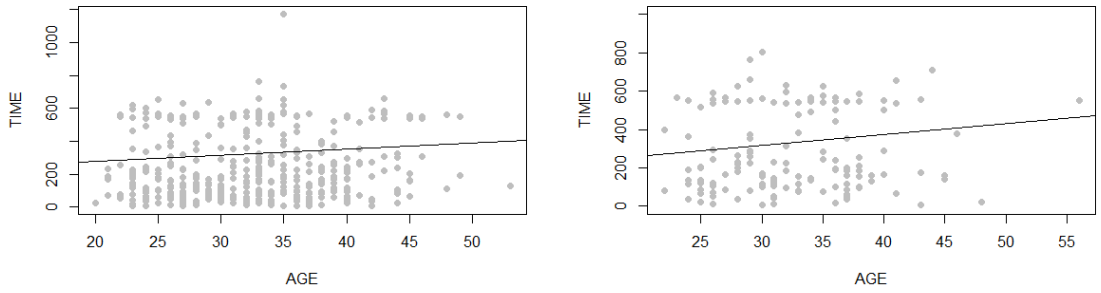
Raza

Realizaremos este estudo separando as razas para comprobar se hai algunha diferencia entre os resultados obtidos para a raza branca e a non branca. Comezaremos analizando con detalle o caso dos/as individuos/as que forman parte do estudo que son de raza branca. Iremos escribindo as funcións empregadas en  e as súas saídas resultantes. Para o caso das persoas de raza non branca podemos observar os resultados obtidos na Táboa 5.2 e unicamente comentaremos que tanto o intercepto como a pendente son significativos e polo tanto son distintos de cero. Realizamos a estimación mediante o método M4, o proposto por Buckley e James, e usaremos a función `bj`:

```
>bj(Surv(TIME,CENSOR)[RACE==0]~AGE[RACE==0],link="identity",x=TRUE, y=TRUE)

Buckley-James Censored Data Regression

bj(formula = Surv(TIME, CENSOR)[RACE == 0] ~ AGE[RACE == 0],
link = "identity", x = TRUE, y = TRUE)

Discrimination
Indexes
Obs    430    Regression d.f.1    g        26.442
Events 357    sigma122.9842
```

```
d.f.      355

           Coef     S.E.    Wald Z   Pr(>|Z|)
Intercept 202.6077  34.8628   5.81   <0.0001
AGE         3.7079  1.0625   3.49    0.0005
```

Nesta saída observamos como tanto o intercepto coma a pendente son significativos pois o nivel crítico de ambos parámetros é menor que os niveis de significación habituais. Obtemos por tanto estimacións de 202.608 días e 3.708 días/anos respectivamente para cada parámetro.



(a) Raza branca.

(b) Raza non branca.

Figura 5.3: Representacións da idade dos/as individuos/as de raza branca e non branca fronte ao tempo ata a recaída en drogas xunto cos axustes obtidos en cada caso.

Se nos fixamos na Táboa 5.2, podemos comprobar que as estimacións teñen o mesmo signo para ambas razas. Polo tanto, podemos observar efectos similares para os dous grupos, onde se obtén unha relación lineal crecente, ao ser a pendente positiva. Entón, podemos concluír que a medida que aumenta a idade dos/as doentes, tanto as persoas de raza branca como de raza non branca van a permanecer máis tempo sen tomar drogas, tal como se observa na Figura 5.3. Ademais, o efecto da idade é máis notable para as persoas de raza non branca xa que a pendente estimada ten un valor máis alto.

Tipo de droga

Realizaremos agora o mesmo procedemento feito para o caso das razas pero separando aos/ás individuos/as dependendo do tipo de droga consumida antes de entrar no ensaio clí-

	Raza	
	Raza 0: Branca	Raza 1: Non branca
$\widehat{\beta}_0$	202.608	146.281
$\widehat{\beta}_1$	3.708	5.724

Táboa 5.2: Táboa das estimacións mediante o método Buckley e James do intercepto e a pendente para o modelo (5.1), considerando a variable explicativa `AGE` con respecto a variable resposta `TIME` diferenciando os resultados obtidos en función da variable `RACE`.

nico: heroína e cocaína, so heroína, so cocaína e ningunha destas dúas drogas. Observamos todas as estimacións obtidas para os diferentes casos na Táboa 5.3. Veremos con detalle a obtención dos parámetros cando o tipo de droga consumida é a cocaína e cando non se consumiu ningunha destas posibles drogas. No caso da cocaína, atopamos que o intercepto non é significativo, polo que teremos que empregar o método de Miller. Destacamos tamén que para o caso de consumir heroína e para ningunha destas drogas, as pendentes non son significativas e polo tanto as variables resposta e explicativa non están relacionadas. Veremos que podemos facer cando nos atopamos esta situación.

	Tipo de droga			
	Heroína e Cocaína	Heroína	Cocaína	Nin cocaína nin heroína
$\widehat{\beta}_0$	-200.285	195.194	-3.790 (NS)	284.161
$\widehat{\beta}_1$	13.406	2.440 (NS)	10.602	-0.169 (NS)

Táboa 5.3: Táboa das estimacións obtidas empregando o método M4 de Buckley e James do intercepto e a pendente no modelo (5.1) cando consideramos a variable explicativa `AGE` con respecto a variable resposta `TIME` diferenciando os resultados obtidos en función da variable `HERCOC`. Nótese que NS significa que dita estimación non resulta estatísticamente significativa.

A continuación, mostramos a saída de  do correspondente estudo cando a droga consumida é a cocaína. Volvemos empregar o método M4 de Buckley e James e así, usamos a función `bj`:

```
>bj(Surv(TIME,CENSOR)[HC==3]~AGE[HC==3],link="identity",x=TRUE, y=TRUE)

Buckley-James Censored Data Regression
```

```
bj(formula = Surv(TIME, CENSOR)[HC == 3] ~ AGE[HC == 3], link = "identity",
x = TRUE, y = TRUE)

Discrimination
Indexes
Obs    172    Regression d.f.1    g       61.561
Events 131    sigma130.2886
d.f.      129

           Coef     S.E.     Wald Z  Pr(>|Z|)
Intercept -3.7902  70.4801   -0.05   0.9571
AGE       10.6016  2.3038     4.60   <0.0001
```

Se nos fixamos no nivel crítico do intercepto, que ten un valor de 0.9571, podemos afirmar que este parámetro non vai ser significativo e polo tanto temos evidencias de que o intercepto é cero. Unha vez concluído isto e usando os coñecementos adquiridos ao longo do Capítulo 4 deste traballo, sabemos que para un modelo sen intercepto o método que mellor estima a pendente é o método M5 proposto por Miller. Polo tanto, usaremos agora a función `CensReg.SMN` para obter unha mellor estimación da pendente:

```
>CensReg.SMN(1-CENSOR[HC==3],AGE[HC==3],TIME[HC==3],cens="right",
dist="Normal")$betas

-----------------------------------------
EM estimates and SE
-----------------------------------------

Estimates          SE
x1          9.69816     0.89852
sigma^2 67560.48197 13478.74770
-----------------------------------------

Model selection criteria
-----------------------------------------

Loglik      AIC      BIC      EDC
Value -958.924 1921.849 1928.144 1923.095
```

```
-----------------------------------------

       [,1]
[1,] 9.698159
```

Obtemos un valor de 9.698 días/anos para a pendente, que é un resultado máis fiable que o obtido mediante o método de Buckley-James que observamos na Táboa 5.3.

Analicemos agora a saída obtida cando as persoas que participaron no ensaio non consumiron nin cocaína nin heroína meses antes de entrar ao tratamento. Empregaremos en primeiro lugar a función `bj`:

```
> bj(Surv(TIME,CENSOR)[HC==4]~AGE[HC==4],link="identity",x=TRUE, y=TRUE)

Buckley-James Censored Data Regression

bj(formula = Surv(TIME, CENSOR)[HC == 4] ~ AGE[HC == 4], link = "identity",
x = TRUE, y = TRUE)

Discrimination
Indexes
Obs    192    Regression d.f.1    g        1.235
Events 152    sigma127.2726
d.f.      150

           Coef     S.E.    Wald Z  Pr(>|Z|)
Intercept 284.1605 52.2488   5.44   <0.0001
AGE        -0.1687  1.5950  -0.11   0.9158
```

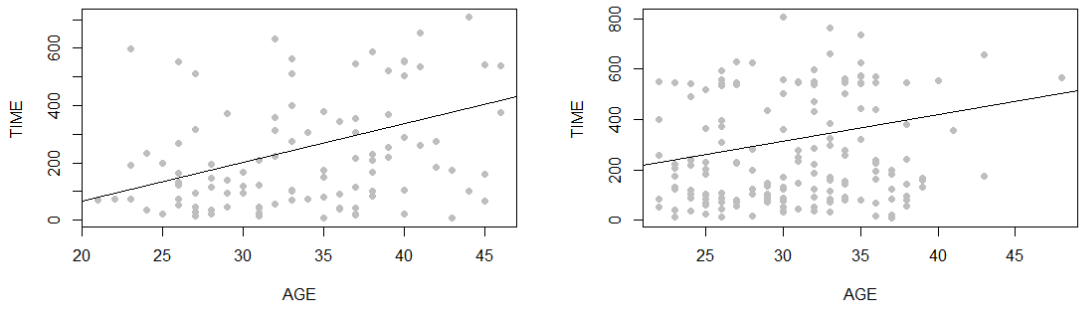
Se observamos o valor do nivel crítico para o intercepto vemos que é significativo. Sen embargo, para o caso da pendente temos un valor de 0.9158, claramente superior aos niveis de significación habituais. Polo tanto podemos concluír que a variable resposta non está relacionada coa variable explicativa cando o/a individuo/a non consume nin cocaína nin heroína meses antes de entrar ao ensaio clínico.

Despois de realizar este estudo, sabemos que a idade das persoas que consumen heroína e ningunha das dúas drogas consideradas non vai estar relacionada co tempo que tardan en recaer no consumo das mesmas. Ademais, tamén sabemos que o intercepto no caso da

cocaína non é significativo e polo tanto podemos empregar o método de Miller que nos da unha mellor estimación da pendente. Con esta información podemos reescribir a Táboa 5.3 con mellores estimacións dando lugar á Táboa 5.4. Podemos ilustrar os datos obtidos nesta táboa na Figura 5.4, onde observamos nestes casos unha relación lineal crecente entre a idade e o tempo de recaída.

	Tipo de droga			
	Heroína e Cocaína	Heroína	Cocaína	Nin cocaína nin heroína
$\widehat{\beta}_0$	-200.285	195.194	0	284.161
$\widehat{\beta}_1$	13.406	0	9.698	0

Táboa 5.4: Táboa das estimacións obtidas empregando o método M5 de Miller do intercepto e a pendente no modelo (5.1) cando consideramos a variable explicativa `AGE` con respecto a variable resposta `TIME` diferenciando os resultados obtidos en función da variable `HERCOC`.



(a) Consumidores/as de heroína e cocaína

(b) Consumidores/as de cocaína

Figura 5.4: Representacións da idade dos/as individuos/as que consumían ou ben heroína e cocaína ou ben so cocaína fronte ao tempo de supervivencia, é dicir, o tempo ata a recaída en drogas.

Trataremos agora o caso no que a variable explicativa non logra explicar a variable resposta no caso de non consumir ningunha das dúas drogas consideradas. Podemos empregar un modelo de regresión lineal múltiple para comprobar se hai outras variables que consigan explicar o modelo. Pese a que neste traballo nos centramos na regresión lineal simple, a efectos prácticos podemos empregar as mesmas funcións empregadas no Capítulo 4 tanto para un modelo simple como para un modelo múltiple. Estudamos as nocións básicas dos

modelos de regresión múltiple na materia de Inferencia Estadística que se pode consultar en [5] e fixemos un breve resumo das características máis importantes na Sección 1.4.

Para isto empregaremos un modelo de regresión lineal múltiple entre a variable resposta `TIME` e as variables explicativas `AGE, BECK, NDT` e `LEN.T` descritas na Táboa 5.1. Recordemos que a variable `AGE` representa a idade do/da doente, `BECK` representa o seu índice de depresión, `NDT` é o número de tratamentos anteriores e finalmente `LEN.T` é o período de estancia no tratamento. Escribimos o modelo da seguinte forma:

$$\texttt{TIME} = \beta_0 + \beta_1 \texttt{AGE} + \beta_2 \texttt{BECK} + \beta_3 \texttt{NDT} + \beta_4 \texttt{LEN.T} + \varepsilon, \tag{5.2}$$

sendo ε o erro que segue unha distribución normal.

Usamos entón a función `bj` e introducimos todas as posibles variables explicativas coas que contamos para observar se algunha variable nos serve para este caso.

```
>bj(Surv(TIME,CENSOR)[HC==4]~AGE[HC==4]+BECK[HC==4]+NDT[HC==4]+LEN.T[HC==4],
link="identity",x=TRUE, y=TRUE)

Buckley-James Censored Data Regression

bj(formula = Surv(TIME, CENSOR)[HC == 4] ~ AGE[HC == 4] + BECK[HC == 4] +
NDT[HC == 4] + LEN.T[HC == 4], link = "identity", x = TRUE, y = TRUE)

Discrimination
Indexes
Obs    192    Regression d.f.4    g      143.170
Events 152    sigma117.2010
d.f.      147

          Coef     S.E.   Wald Z   Pr(>|Z|)
Intercept 179.5252 54.4496   3.30   0.0010
AGE        -0.6428  1.5267  -0.42   0.6737
BECK       -0.3196  1.1221  -0.28   0.7757
NDT        -8.1776  2.0986  -3.90   <0.0001
LEN.T       1.6183  0.1332  12.15   <0.0001
```

Á vista dos resultados, observamos que tanto a idade como o índice de depresión non son variables que inflúan no tempo ata a recaída nas drogas posto que os seus niveis críticos

teñen valores superiores aos niveis de significación habituais e polo tanto as súas pendentes non van ser significativamente distintas de cero.

Por outra banda, observamos que o número de tratamentos anteriores (`NDT`) e o número de días que permanecen no ensaio (`LEN.T`) si que están relacionados coa variable resposta `TIME`. En canto ao número de tratamentos anteriores, tendo en conta que a estimación da pendente é negativa, teremos un efecto decrecente en canto ao número de días antes da recaída sempre que as demais variables se manteñan constantes. Observamos que ocorre o contrario co número de días que un/unha individuo/a permanece no ensaio xa que a estimación da pendente neste caso é positiva e polo tanto comprobamos que a variable resposta aumenta ao aumentar dita variable explicativa, sempre que as demais variables se manteñan constantes.

Capítulo 6

Conclusións

Os datos censurados xorden de maneira natural cando existe unha limitación na información que dispoñemos dunha certa variable de interese. Ao longo deste traballo centrámonos na censura aleatoria pola dereita aínda que tamén hai outros tipos tal como se estudou na Sección 2.2. Neste contexto, en lugar de empregar a clásica función de distribución empírica como se fai no caso de datos completos, empregaremos a función de Supervivencia e a función de risco para tratar de resumir este tipo de datos tal e como se expón na Sección 2.3.

Cando temos datos censurados, non podemos empregar as mesmas técnicas e métodos que no caso de observar os datos completos, como puidemos comprobar durante a realización deste traballo. No Capítulo 3 estudamos diferentes métodos para estimar os coeficientes asociados a un modelo de regresión con variable resposta censurada pola dereita como os métodos propostos por Stute, Miller, Buckley e James ou Jin, Lin e Ying. En primeiro lugar, no método de Stute cobran vital importancia os pesos Kaplan-Meier estudados na Sección 2.5. No método de Miller empregamos o método de máxima verosimilitude e obtemos a solución grazas ao método de Newton-Rapson. En canto ao método proposto por Buckley e James, destacamos o uso do método de mínimos cadrados usual pero realizando algunha variante para o caso dos datos censurados. Finalmente, o método de Jin, Lin e Ying intenta mellorar o estimador de Buckley e James para solucionar os seus problemas de converxencia.

No Capítulo 4 realizamos un estudo de simulación que nos permitiu comparar os distintos estimadores estudados ao longo do Capítulo 3. Ademais, diferenciamos entre modelos con intercepto e modelos sen intercepto, obtendo unhas conclusións diferentes en cada caso. En primeiro lugar, no caso do modelo con intercepto, o método que menores erros cadrá-

ticos medios proporcionou foi o método de Buckley e James, ao que nos referimos como M4 durante todo o traballo. Por outro lado, no caso do modelo sen intercepto, observamos que o mellor método é o de Miller, M5. Debemos destacar que, en ambos casos, o segundo mellor método foi o método de Stute, denotado ao longo do traballo por M2. Neste estudo de simulación, tamén tivemos en conta a distribución do erro e diferenciamos o caso no que o erro ten unha distribución normal e o caso no que ten unha distribución chi-cadrado, obtendo peores resultados neste último caso tal e como esperabamos posto que os estimadores considerados foron deseñados baixo o suposto de normalidade do erro. Sen embargo, o método proposto por Buckley e James seguiu proporcionando unhas boas estimacións pese á distribución do erro.

Finalmente, no Capítulo 5, empregamos a base de datos `UIS` de  para ilustrar os métodos estudados nun caso real. Esta base de datos contén os datos asociados a un ensaio clínico deseñado para analizar diferentes tratamentos para combater a adicción ao uso de drogas. Consideramos como variable de interese (variable resposta) o tempo que pasa dende que un/unha doente entra no ensaio ata que sofre unha recaída nas drogas. Estudamos se hai algunha relación entre dita variable e a idade dos/das doentes e chegamos a que teñen unha relación crecente para o caso máis xeral. Logo estudamos varios subgrupos grazas ás variables categóricas ofrecidas pola base de datos empregada e observamos, por exemplo, que a idade dos/as individuos/as que consumían cocaína antes de entrar ao tratamento, non inflúe na variable resposta que recolle o tempo ata a recaída.

Anexo A: Comandos de R

A.1. Figuras do Capítulo 2

Exemplo datos censurados, Figura 2.1

```
x=0
y=0
plot(x,y,col="white")
segments(1995, 5, 8,5)

plot(x,y,pch=16,col="black",
xlab="Tempo",ylab="Doentes",xlim=c(2000,2010),ylim=c(6,1))

segments(2000, 1, 2008,1)
segments(2000, 2, 2006,2)
segments(2001, 3, 2009,3)
segments(2002, 4, 2008,4)
segments(2002, 5, 2004,5)
segments(2002, 6, 2006,6)

points(2007,1)
points(2008,1,pch=4)
points(2006,2, pch=4)
points(2007,3)
points(2009,3,pch=4)
points(2007,4)
points(2008,4, pch=4)
points(2004,5,pch=4)
```

```
points(2006,6,pch=4)

points(2000,1, pch=16)
points(2000,2, pch=16)
points(2001,3, pch=16)
points(2002,4, pch=16)
points(2002,5, pch=16)
points(2002,6, pch=16)

abline(v=2000,lty=3,col="black")
abline(v=2002.5,lty=3,col="black")
abline(v=2007,lty=3,col="black")
```

Gráfica motivación datos censurados, Figura 2.2

```
#Tamaño n=100
n=100

xseq=seq(-3,3,by=0.01)
plot(xseq,pnorm(xseq),type="l",lwd=2,xlab="y",ylab="Distribución")
legend(x = "bottomright", legend = c("Datos censurados",
 "Datos completos"), fill = c("blue", "red"))

x=rnorm(n)
lines(ecdf(x),col="red",lwd=2)

xc=x[x<=1]
lines(ecdf(xc),col="blue",lwd=2)

#Para tamaño n=10000
n=1000

xseq=seq(-3,3,by=0.01)
plot(xseq,pnorm(xseq),type="l",lwd=2,xlab="y",ylab="Distribución")
legend(x = "bottomright", legend = c("Datos censurados",
"Datos completos"),fill = c("blue", "red"))
```

```
x=rnorm(n)
lines(ecdf(x),col="red",lwd=6)

xc=x[x<=1]
lines(ecdf(xc),col="blue",lwd=2)
```

Gráfica Kaplan-Meier, Figura 2.3

```
n=10
x=runif(n,min=0,max=1)
c=rnorm(n)+4
error=rnorm(n,sd=0.5)
y=theta[1]+theta[2]*x+error
z=pmin(y,c)
delta=as.numeric(y<=c)

datcens <- data.frame(t = x, cen = delta)

library(survival)
fit <- survfit(Surv(t, cen)~1, data = datcens)
summary(fit)

plot(fit, main = "Metodo de Kaplan-Meier")
legend("bottomleft",  c("F. Supervivencia", "Intervalo de confianza"),
lty = 1:2)
```

A.2. Comandos do Capítulo 4

Gráfico comparativo de varianzas, Figura 4.1

```
#Varianza=0.5
n=100
theta=c(1,2)
x=runif(n,min=0,max=1)
```

```
c=rnorm(n)+2.8# variable censurada
error=rnorm(n,sd=0.5) # Error epsilon
y=theta[1]+theta[2]*x+error # variable resposta
z=pmin(y,c)
delta=as.numeric(y<=c)

plot(x,y,col=8,type="p",pch=19, xlab="x", ylab="y")
abline(1,2) #valores reales

#varianza=1
n=100
x=runif(n,min=0,max=1)
c=rnorm(n)+3# variable censurada
error=rnorm(n,sd=1) # Error epsilon
y=theta[1]+theta[2]*x+error # variable resposta
z=pmin(y,c)
delta=as.numeric(y<=c)

plot(x,y,col=8,type="p",pch=19,xlab="x", ylab="y")
abline(1,2) #valores reales
```

Gráfico da Figura 4.2

```
n=100
theta=c(1,2)
x=runif(n,min=0,max=1)
c=rnorm(n)+2.8# variable censurada
error<-rchisq(n,df=3) # Error Xi cadrado
y=theta[1]+theta[2]*x+error # variable resposta
z=pmin(y,c)

plot(x,y,col=8,type="p",pch=19, xlab="x", ylab="y")
abline(1,2) #valores reales
```

Simulación do Modelo 4.2.1

```
#MODELO CON INTEREPTO (MODELO 1)
```

```
#CENSURA DUN 25%

#Cargamos as librerias que usaremos
library(survival)
library(condSURV) #Librería para os pesos Kaplan-Meier
library(xtable) #Librería para as táboas dos datos
library(rms) #Librería para Buclkey-James

##############-------------------------------------------------

#CASO DE DESVIACIÓN TÍPICA DO ERRO sd=0.5
#O resto de casos son análogos

##############-------------------------------------------------

#Calculamos a censura:

n=10000
x=runif(n,min=0,max=1)
c=rnorm(n)+2.84 # Variable censurada
error=rnorm(n,sd=0.5) # Error epsilon
y=1+2*x+error # variable resposta
z=pmin(y,c)
delta=as.numeric(y<=c)
censura=100-sum(delta)/length(delta)*100 #Porcentaxe de censura
censura

############-----------------------------------------------------

set.seed(123456) #Fixamos semilla

#Tamaño de mostra
n=100
#n=500
#n=1000
```

```
m=1000 #Número de simulacións
theta=c(1,2) #O beta teórico

# As matrices theta_lm, theta_KM, theta_BJ e theta_Jin teran m filas
# e 2 columnas: a primeira dos alphas (beta_0) e a segunda dos beta_1
theta_lm=matrix(nrow = m,ncol=2)
theta_Stute=matrix(nrow=m,ncol=2)
theta_KMS=matrix(nrow=m,ncol=2)
theta_BJ=matrix(nrow=m,ncol=2)
fallou=numeric(m) #vector empregado en Buclkey-James(M4)

#Gardaremos os datos resultantes nunha matriz que defino agora:
M=matrix(NA,nrow=4,ncol=6)

for (i in 1:m){

x=runif(n,min=0,max=1)
c=rnorm(n)+2.84 # Variable censurada
error=rnorm(n,sd=0.1) # Error epsilon
y=theta[1]+theta[2]*x+error # variable resposta
z=pmin(y,c)
delta=as.numeric(y<=c)
indices=sort(z,index.return=T)$ix #Ordeno os datos
z=z[indices]
x=x[indices]
delta=delta[indices]
delta[n]=1 #Defino o ultimo dato como non censurado

#Facemos agora os diferentes metodos:

#MÉTODO M1:mínimos cadrados usual para datos sen censura
#Escollemos os z e x que cumpran delta=1 (datos sen censura)
lm(z[delta==1]~x[delta==1])
#O vector coef contén os alpha e beta estimados coa función lm
coef=coefficients(lm(z[delta==1]~x[delta==1]))
#Gardo na primeira columna os alphas e na segunda os betas
```

```
theta_lm[i,1:2]=coef

#MÉTODO M2: Stute
peso_KM=KMW(z,delta) #Calculo os pesos KM coa función KMW
#Calculo os coeficientes empregando a función lm cos pesos KM
coefStute=coefficients((lm(z~x,weights = peso_KM)))
theta_Stute[i,1:2]=coefStute

#MÉTODO M3: Stute con pesos KMS
peso_KMS=PKMW(z,delta) #Calculo os pesos KM Suavizados coa función
PKMW
peso_KMS[n]=peso_KMS[n]+1-sum(peso_KMS)
#Calculo os coeficientes empregando a función lm cos pesos KMS
coefKMS=coefficients((lm(z~x,weights = peso_KMS)))
theta_KMS[i,1:2]=coefKMS

#MÉTODO M4: Buclkey-James
#Empregamos a función bj
modeloBJ=bj(Surv(z,delta)~x,link="identity",x=TRUE, y=TRUE)
if (length(names(modeloBJ)) !=20){
fallou[i]=1
} else{

coefBJ=modeloBJ$coefficients
theta_BJ[i,1:2]=coefBJ
}
}

#Gardamos os datos nun arquivo:
datos=cbind(theta_lm,theta_Stute,theta_KMS,theta_BJ)
colnames(datos)=c("M1_I","M1_P","M2_I","M2_P", "M3_I", "M3_P", "M4_I",
"M4_P")

#Para n=100
write.table(datos, file = "modelo1_c25_100A.txt", sep = " " ,quote=F,
col.names = colnames(datos),row.names=FALSE)
```

```
#Para n=500
#write.table(datos, file = "modelo1_c25_500A.txt", sep = " " ,quote=F,
col.names = colnames(datos),row.names=FALSE)
#Para n=1000
#write.table(datos, file = "modelo1_c25_1000A.txt", sep = " " ,quote=F,
col.names = colnames(datos),row.names=FALSE)

#Comprobamos que os escribimos correctamente:
proba=read.table("modelo1_c25_100A.txt",header=TRUE)
head(proba)

#Calculo sesgo, varianza e ECM para cada caso

#MÉTODO M1
media_lm=colMeans(theta_lm)
sesgo_lm=media_lm-theta
varianza_lm=c(var(theta_lm[,1]),var(theta_lm[,2]))
ECM_lm=sesgo_lm^2+varianza_lm
l1=c(sesgo_lm[1],varianza_lm[1],ECM_lm[1]) #correspondente a beta_0
l2=c(sesgo_lm[2],varianza_lm[2],ECM_lm[2]) #correspondente a beta_1
M[1,1:6]=c(l1,l2)

#MÉTODO M2
media_Stute=colMeans(theta_Stute)
sesgo_Stute=(media_Stute-theta)
varianza_Stute=c(var(theta_Stute[,1]),var(theta_Stute[,2]))
ECM_Stute=sesgo_Stute^2 +varianza_Stute

Stute1=c(sesgo_Stute[1],varianza_Stute[1],ECM_Stute[1]) #beta_0
Stute2=c(sesgo_Stute[2],varianza_Stute[2],ECM_Stute[2]) #beta_1
M[2,1:6]=c(Stute1,Stute2)

#MÉTODO M3
media_KMS=colMeans(theta_KMS)
sesgo_KMS=media_KMS-theta
varianza_KMS=c(var(theta_KMS[,1]),var(theta_KMS[,2]))
```

```
ECM_KMS=sesgo_KMS^2 +varianza_KMS
KMS1=c(sesgo_KMS[1],varianza_KMS[1],ECM_KMS[1]) #beta_0
KMS2=c(sesgo_KMS[2],varianza_KMS[2],ECM_KMS[2]) #beta_1
M[3,1:6]=c(KMS1,KMS2)

#MÉTODO M4
media_BJ=colMeans(theta_BJ[which(fallou !=1),])
sesgo_BJ=media_BJ-theta
varianza_BJ=c(var(theta_BJ[which(fallou !=1),1]),
var(theta_BJ[which(fallou !=1),2]))
ECM_BJ=sesgo_BJ^2 +varianza_BJ
BJ1=c(sesgo_BJ[1],varianza_BJ[1],ECM_BJ[1]) #beta_0
BJ2=c(sesgo_BJ[2],varianza_BJ[2],ECM_BJ[2]) #beta_1
M[4,1:6]=c(BJ1,BJ2)

#Gardaremos os datos resultantes nunha matriz

rownames(M)<-c("M1","M2", "M3", "M4")
colnames(M)<-c("Sesgo", "Var","ECM","Sesgo", "Var","ECM")
M
M*10000
#Obtemos o sesgo, varianza e ECM de cada un dos metodos multiplicados
por 10000

xtable(M*10000,digits = 3) #Para obter os comandos de Latex
```

Simulación do Modelo 4.3.1

```
#MODELO SEN INTERCEPTO (MODELO 2)


#CENSURA DUN 50%

#Cargamos as librerias que usaremos
library(survival)
library(condSURV) #Libreria para os pesos Kaplan-Meier
library(SMNCensReg) #Librería para o método de Miller
```

```
library(xtable) #Libreria para as táboas dos datos
library(rms) #Libreria para o método de Buckley-James
library(lss2) #Libreria para o método de Jin, Lin e Ying

##############-------------------------------------------------

#CASO DE DESVIACIÓN TÍPICA DO ERRO: sd=0.5
#O resto de casos son análogos

##############-------------------------------------------------

#Calculamos a censura:

n=10000
x=runif(n,min=0,max=1)
c=rnorm(n)+1 # Variable censurada
error=rnorm(n,sd=0.5) # Error epsilon
y=2*x+error # variable resposta
z=pmin(y,c)
delta=as.numeric(y<=c)
censura=100-sum(delta)/length(delta)*100 #Porcentaxe de censura
censura

############-----------------------------------------------------

set.seed(123456) #Fixamos semilla

#Tamaño da mostra
#n=50
#n=100
n=200

m=1000 #Numero de simulacións

theta=2 #O beta teórico
```

```
# Os vectores theta_lm, theta_Stute, theta_KMS, theta_Miller,
# theta_BJ e theta_Jin son vectores de m compoñentes onde
# iremos gardando os resultados da estimación

theta_lm=c(rep(NA,m))
theta_Stute=c(rep(NA,m))
theta_KMS=c(rep(NA,m))
theta_Miller=c(rep(NA,m))
theta_BJ=c(rep(NA,m))
theta_Jin=c(rep(NA,m))
fallou=numeric(m) #vector empregado en Buckley-James

#Gardaremos os datos resultantes nunha matriz que defino agora:
M=matrix(NA,nrow=6,ncol=3)

for (i in 1:m){
set.seed(i)

x=runif(n,min=0,max=1)
c=rnorm(n)+1 # Variable censurada (para unha censura dun 25%)
error=rnorm(n,sd=0.5) #Error epsilon
y=theta*x+error #variable resposta
z=pmin(y,c)
delta=as.numeric(y<=c)

#Ordeno os datos
indices=sort(z,index.return=T)$ix
z=z[indices]
x=x[indices]
delta=delta[indices]

#Defino o ultimo dato como non censurado
delta[n]=1

#Fago agora os diferentes metodos:
```

```
#MÉTODO M1: mínimos cadrados usual para datos sen censura
#Escollemos os z e x que cumpran delta=1 ( datos sen censura)
lm(z[delta==1]~x[delta==1]-1)
#O vector coef contén o beta estimado coa función lm
theta_lm[i]=coefficients(lm(z[delta==1]~x[delta==1]-1))
theta_lm[i]

#MÉTODO M2: Stute
#Calculo os pesos KM coa función KMW
peso_KM=KMW(z,delta)
#Calculo os coeficientes empregando a función lm cos pesos KM
theta_Stute[i]=coefficients((lm(z~x-1,weights = peso_KM)))

#MÉTODO M3: Stute con pesos KMS
#Calculo os pesos KM Suavizados coa función PKMW
peso_KMS=PKMW(z,delta)
peso_KMS[n]=peso_KMS[n]+1-sum(peso_KMS)
#Calculo os coeficientes empregando a función lm cos pesos KMS
theta_KMS[i]=coefficients((lm(z~x-1,weights = peso_KMS)))

#MÉTODO M4: Buckley-James
#Empregamos a función bj
modeloBJ=bj(Surv(z,delta)~x,link="identity",x=TRUE, y=TRUE)
if (length(names(modeloBJ)) !=20){
fallou[i]=1
} else{
theta_BJ[i]=modeloBJ$coefficients[2]
}

#MÉTODO M5: Miller
modelo_Miller=CensReg.SMN(1-delta,x,z,cens="right",dist="Normal")
theta_Miller[i]=modelo_Miller$betas

#MÉTODO M6: Jin, Lin e Ying
modeloJin=lss(Surv(z,delta)~x,maxiter = 3)
theta_Jin[i]=as.numeric(modeloJin$lse)
```

```
cat(paste("\n \n Estamos en la iteracion ",i,".\n",sep=""),
collapse="\n")
}

#Gardamos os datos nun arquivo:
datos=cbind(theta_lm,theta_Stute,theta_KMS,theta_Miller,
theta_BJ, theta_Jin)
colnames(datos)=c("M1","M2", "M3", "M4", "M5", "M6")

#n=50
#write.table(datos, file = "modelo2_c50_50A.txt", sep = " " ,quote=F,
col.names = colnames(datos),row.names=FALSE)
#n=100
#write.table(datos, file = "modelo2_c50_100A.txt", sep = " " ,quote=F,
col.names = colnames(datos),row.names=FALSE)
#n=200
write.table(datos, file = "modelo2_c50_200A.txt", sep = " " ,quote=F,
col.names = colnames(datos),row.names=FALSE)

#Calculo sesgo, varianza e ECM para cada caso:

#MÉTODO M1
media_lm=mean(theta_lm)
sesgo_lm=media_lm-theta
varianza_lm=var(theta_lm)
ECM_lm=sesgo_lm^2+varianza_lm
M[1,1:3]=c(sesgo_lm,varianza_lm,ECM_lm)

#MÉTODO M2
media_Stute=mean(theta_Stute)
sesgo_Stute=media_Stute-theta
varianza_Stute=var(theta_Stute)
ECM_Stute=sesgo_Stute^2 +varianza_Stute
M[2,1:3]=c(sesgo_Stute,varianza_Stute,ECM_Stute)
```

```
#MÉTODO M3
media_KMS=mean(theta_KMS)
sesgo_KMS=media_KMS-theta
varianza_KMS=var(theta_KMS)
ECM_KMS=sesgo_KMS^2 +varianza_KMS
M[3,1:3]=c(sesgo_KMS,varianza_KMS,ECM_KMS)

#MÉTODO M4
media_BJ=mean(theta_BJ[which(fallou !=1)])
sesgo_BJ=media_BJ-theta
varianza_BJ=var(theta_BJ[which(fallou !=1)])
ECM_BJ=sesgo_BJ^2 +varianza_BJ
M[4,1:3]=c(sesgo_BJ,varianza_BJ,ECM_BJ)

#MÉTODO M5
media_Miller=mean(theta_Miller)
sesgo_Miller=media_Miller-theta
varianza_Miller=var(theta_Miller)
ECM_Miller=sesgo_Miller^2 +varianza_Miller
M[5,1:3]=c(sesgo_Miller,varianza_Miller,ECM_Miller)

#MÉTODO M6
media_Jin=mean(theta_Jin)
sesgo_Jin=media_Jin-theta
varianza_Jin=var(theta_Jin)
ECM_Jin=sesgo_Jin^2 +varianza_Jin
M[6,1:3]=c(sesgo_Jin,varianza_Jin,ECM_Jin)

rownames(M)<-c("M1","M2", "M3", "M4", "M5","M6")
colnames(M)<-c("Sesgo", "Var","ECM")
#Obtemos o sesgo, varianza e ECM de cada un dos metodos
M

xtable(M,digits = 3)
```

A.3. Comandos do Capítulo 5

Histogramas: Figura 5.1

```
#Cargamos a librería para a base de datos uis
library(quantreg)
data(uis)
datos=uis
head(uis) #Observamos os 10 primeiros valores das variables
attach(datos) #Para poder usar os nomes dos datos
summary(datos)

#Cálculo da censura
censura=100-sum(CENSOR)/length(CENSOR)*100 #Porcentaxe de censura
censura

#---------------------------------------------------------
#---------------ESTUDO DOS DATOS-------------------------
#Realizaremos uns gráficos iniciais

hist(AGE, main = "",
xlab = "Idade", ylab = "Frecuencia",col = "ivory")

hist(BECK,main="",
xlab = "Depresión",
ylab = "Frecuencia",col = "ivory")

hist(NDT,main="",
xlab = "Número de tratamentos anteriores",
ylab = "Frecuencia",col = "ivory")

hist(LEN.T,main="", xlab = "Número de días no tratamento",
ylab = "Frecuencia",col = "ivory")
```

Figura 5.2

```
plot(AGE,TIME,col="gray",pch=16,xlab="AGE", ylab="TIME")
abline(bj(Surv(TIME,CENSOR)~AGE,link="identity",x=TRUE, y=TRUE))
```

Figura 5.3

```
#RAZA BRANCA
plot(AGE[RACE==0],TIME[RACE==0],col="gray",pch=16,xlab="AGE",
ylab="TIME")
abline(bj(Surv(TIME,CENSOR)[RACE==0]~AGE[RACE==0],link="identity",
x=TRUE, y=TRUE))

#RAZA NON BRANCA
plot(AGE[RACE==1],TIME[RACE==1],col="gray",pch=16,xlab="AGE",
ylab="TIME",ylim=c(0,1000))
abline(bj(Surv(TIME,CENSOR)[RACE==1]~AGE[RACE==1],link="identity",
x=TRUE, y=TRUE))
```

Análise da base de datos reais

```
#Cargamos as librerías necesaria
library(quantreg) #Librería para a base de datos uis
library(survival)
library(condSURV) #Librería para os pesos Kaplan-Meier
library(xtable) #Librería para as táboas dos datos
library(rms) #Librería para Buclkey-James
library(SMNCensReg)

#-------------------------------------------------
#Cargamos os datos:
data(uis)
datos=uis
head(uis) #Observamos os 10 primeiros valores das variables
attach(datos) #Para poder usar os nomes dos datos
summary(datos)
```

```
# Estudaremos a relación entre a variable idade, AGE, e a variable
# resposta TIME. Consideraremos varios subgrupos coas variables
3 categóricas HERCOC e RACE.


#-------------------------------------------------
#VARIABLE HERCOC
#-------------------------------------------------

#-------HC=1:Consumo de heroína e cocaína---------

bj(Surv(TIME,CENSOR)[HC==1]~AGE[HC==1],link="identity",x=TRUE,
y=TRUE)
plot(AGE[HC==1],TIME[HC==1])
abline(bj(Surv(TIME,CENSOR)[HC==1]~AGE[HC==1],link="identity",
x=TRUE, y=TRUE))

#---------------------HC=2:Heroina-------------------

bj(Surv(TIME,CENSOR)[HC==2]~AGE[HC==2],link="identity",x=TRUE,
y=TRUE)

#Modelo de regresión múltiple

bj(Surv(TIME,CENSOR)[HC==2]~AGE[HC==2]+BECK[HC==2]+NDT[HC==2]+
LEN.T[HC==2],
link="identity",x=TRUE, y=TRUE)

#---------------------HC=3:Cocacaina-------------------

bj(Surv(TIME,CENSOR)[HC==3]~AGE[HC==3],link="identity",x=TRUE,
y=TRUE)
plot(AGE[HC==3],TIME[HC==3])
abline(bj(Surv(TIME,CENSOR)[HC==3]~AGE[HC==3],link="identity",
x=TRUE, y=TRUE))
```

```
# Intercepto non significativo, entón usaremos o método de Miller

CensReg.SMN(1-CENSOR[HC==3],AGE[HC==3],TIME[HC==3],cens="right",
dist="Normal")$betas

#------------------HC=4:Ningunha destas drogas-----------------

bj(Surv(TIME,CENSOR)[HC==4]~AGE[HC==4],link="identity",x=TRUE,
y=TRUE)

bj(Surv(TIME,CENSOR)[HC==4]~AGE[HC==4]+BECK[HC==4]+NDT[HC==4]+
LEN.T[HC==4],
link="identity",x=TRUE, y=TRUE)

#-------------------------------------------------
#RACE
#-------------------------------------------------

#---------------------RACE 0: branca-------------------

bj(Surv(TIME,CENSOR)[RACE==0]~AGE[RACE==0],link="identity",x=TRUE,
y=TRUE)
plot(AGE[RACE==0],TIME[RACE==0])
abline(bj(Surv(TIME,CENSOR)[RACE==0]~AGE[RACE==0],link="identity",
x=TRUE, y=TRUE))

#--------------------RACE 1: non branca-----------------

bj(Surv(TIME,CENSOR)[RACE==1]~AGE[RACE==1],link="identity",x=TRUE,
y=TRUE)
plot(AGE[RACE==1],TIME[RACE==1])
abline(bj(Surv(TIME,CENSOR)[RACE==1]~AGE[RACE==1],link="identity",
x=TRUE, y=TRUE))
```

Bibliografía

[1] Buckley, J. e James, I., *Linear regression with censored data*, Biometrika, **66** (1979), 429–436.

[2] Cao R., López-de-Ullibarri, I., Janssen, P. e Veraverbeke, N., *Presmoothed Kaplan-Meier and Nelson-Aalen estimators*, Journal of Nonparametric Statistics, **17** (2005), 31–56.

[3] Dikta, G., *On semiparametric random censorship models*, Journal of Statistical Planning and Inference, **66** (1998), 253–279.

[4] Gehan, E.A., *A generalized Wilcoxon test for comparing arbitrarily singlecensored samples*, Biometrika, **52** (1965), 203—223.

[5] Gómez Villegas, M.A., *Inferencia Estadística*, Díaz de Santos (2005).

[6] Hosmer, D. W., Lemeshow, S. e May, S., *Applied Survival Analysis, Regression Modeling of Time-to-Event Data*, Wiley, 2nd ed., (2008).

[7] Jin, Z., Lin, D. Y., Wei, L. J e Ying, Z., *Rank-based inference for the accelerated failure time model*, Biometrika, **90** (2003), 341–53.

[8] Jin, Z., Lin, D. Y. e Ying, Z., *On least-squares regression with censored data*, Biometrika, **93** (2006), 147–161.

[9] Kaplan, E. L. e Meier, P., *Nonparametric estimation from incomplete observations*, Journal of the American Statistical Association, **53** (1958), 457–481

[10] Lai, T. L. e Ying, Z., *Large sample theory of modified Buckley-James estimator for regression analysis with censored data*, Annals of Statistics, **10** (1991), 1370–402.

[11] Miller, R., Gong, G. e Muñoz, A., *Survival Analysis*, Thechnical report, Universidade de California (1980).

[12] Moore, D. F., *Applied Survival Analysis Using R*, Springer (2016).

[13] Ritov, Y., *Estimation in a linear regression model with censored data*, Annals of Statistics, **18** (1990), 303–28.

[14] Sánchez Sellero C., *Inferencia Estadística en datos con censura y/o truncamiento*, Universidade de Santiago de Compostela (2001). Tese de doutoramento.

[15] Sánchez Sellero, C., *Apuntes da materia Inferencia Estatística*, Grao en Matemáticas, Universidade de Santiago de Compostela, 2018-2019.

[16] Stute, W., *Nonlinear censored regression*, Statistica Sinica, **9** (1999), 1089–1102.

[17] Stute, W., González Manteiga, W. e Sánchez Sellero, C., *Nonparametric Model Checks In Censored Regression*, Communication in Statistics-Theory and Methods, **29** (2000), 1611–1629.